Aspect-based Key Point Analysis for Quantitative Summarization of Reviews

An Quang Tang
RMIT University, Australia
s3695273@rmit.edu.vn

Xiuzhen Zhang
RMIT University, Australia
xiuzhen.zhang@rmit.edu.au

Minh Ngoc Dinh
RMIT University, Australia
minh.dinh4@rmit.edu.vn

Abstract

Key Point Analysis (KPA) is originally for summarizing arguments, where short sentences containing salient viewpoints are extracted as key points (KPs) and quantified for their prevalence as salience scores. Recently, KPA was applied to summarize reviews, but the study still relies on sentence-based KP extraction and matching, which leads to two issues: sentence-based extraction can result in KPs of overlapping opinions on the same aspects, and sentence-based matching of KP to review comment can be inaccurate, resulting in inaccurate salience scores. To address the above issues, in this paper, we propose Aspect-based Key Point Analysis (ABKPA), a novel framework for quantitative review summarization. Leveraging the readily available aspect-based sentiment analysis (ABSA) resources of reviews to automatically annotate silver labels for matching aspect-sentiment pairs, we propose a contrastive learning model to effectively match KPs to reviews and quantify KPs at the aspect level. Especially, the framework ensures extracting KP of distinct aspects and opinions, leading to more accurate opinion quantification. Experiments on five business categories of the popular Yelp review dataset show that ABKPA outperforms state-of-the-art baselines. Source code and data are available at: `https://github.com/antangrocket1312/ABKPA`

1 Introduction

Summarization of user reviews on the online marketplace has become essential both for businesses to improve their product and service qualities and for customers to make purchasing decisions. Although the star ratings aggregated from customer reviews are widely used to measure quality of service for business entities (McGlohon et al., 2010; Tay et al., 2020), they can not explain specific details to achieve business inteligence and informed decision. Early studies on review summarization focus on textual summaries that only represent the major opinions in reviews (Dash et al., 2019; Shandilya et al., 2018) but ignore the minority opinions and fail to quantify the opinion prevalence.

Recently, the quantitative view was introduced to review summarization under the novel framework named Key Point Analysis (KPA) (Bar-Haim et al., 2020a,b, 2021). KPA studies were initially extractive and developed for argument summarization (Bar-Haim et al., 2020a,b), and are then adapted for business reviews (Bar-Haim et al., 2021). KPA consists of two subtasks, namely Key Point extraction, which extracts salient sentences as KPs, and Key Point Matching, which quantifies the prevalence of KPs as the number of matching comments in reviews [1]. More recent KPA studies used abstractive summarization models to generate salient KPs (Kapadnis et al., 2021; Li et al., 2023a).

Whether extractive or abstractive approaches, existing KPA studies still perform KP extraction and matching at the sentence level, which has two major issues. First, the extracted KPs (i.e., short sentences) can contain overlapping opinions on the same aspects, causing high KP redundancy. Subsequently, with both comments and KPs containing multiple opinions, sentence-based matching of KPs to comment then becomes ineffective and results in inaccrurate KP prevalence.

To address the two above issues, we propose Aspect-based Key Point Analysis (ABKPA), a novel and more effective extractive KPA framework for review summarization. ABKPA comprises two key components: Aspect-based KP extraction and Aspect-based KP Matching. First, leveraging the fine-grained aspect-based sentiment analysis (ABSA) model (Miao et al., 2020) for review comments, ABKPA extracts KPs free from redundancy and containing single opinions. Next, again making use of readily available ABSA re-

[1] A comment is a senence in reviews

Findings of the Association for Computational Linguistics: EACL 2024, pages 1419–1433
March 17-22, 2024 ©2024 Association for Computational Linguistics

Table 1: An example showing the summary output of ABKPA and sentence-based KPA (Bar-Haim et al., 2021). Given (a) The input comments, we exemplify and compare the output of (b) sentence-based KPA and (c) ABKPA. In (b) and (c), the columns "Matched comments" and "Quantity" illustrate matching KPs to comments and quantifying KPs in the summary.

(a) **The input comments**. Each box represents a review containing several comments

Review	**Comments (review sentences)**
1	**1.1:** The service is great and the staff is friendly and engaging. **1.2:** The food is excellent but the portion is quite small and quite expensive.
2	**2.1:** The food has great taste but very small portion and the service is slow.
3	**3.1:** The service was good and the food was delicious. **3.2:** Staff is friendly and attentive.
4	**4.1:** Food was excellent and delicious. **4.2:** Service and staff are excellent.
...	...

(b) Sentence-based KPs and their salience score (Bar-Haim et al., 2021, 2020a) output. Note that a commment can only be matched with one KP on of highest confidence.

Key points	**Matched Comments**	**Salience score**
KP1: Service and staff are excellent.	1.1	1
KP2: Service was prompt and friendly. (*redundant*)	3.1	1
...	...	...
KP3: Small and overpriced portion.	1.2	1
KP4: Small food portion and slow service. (*redundant*)	2.1	1
...	...	...

(c) **ABKPA** KPs and their salience score. ABKPA ensures retrieving single-aspect key points with better opinion quantification specific to every comment's aspect

Key points	**Matched Comments**	**Salience score**
KP1: Food was excellent and delicious.	1.2; 2.1; 3.1	3
KP2: Service was prompt and friendly.	1.1; 3.1	2
KP3: Staff is friendly and attentive.	1.1	1
...	...	...
KP4: Small and overpriced portion.	1.2; 2.1	2
KP5: Service was poor and slow	2.1	1
...	...	...

sources for automatic annotation of silver labels for matching aspect-sentiment pairs, we design a contrastive learning model to learn a better representation of opinions in KPs and comments, which provides more a accurate salient score of KPs for better opinion quantification.

Table 1 presents a comparison between ABKPA and sentence-based KPA (Bar-Haim et al., 2020a, 2021). As an example, consider the long comment "2.1: The food has great taste but very small portion and the service is slow.". In Table 1b, sentence-based KPA, applying the supervised matching model from the argument domain at the sentence level, can only match this comment to *one* KP "KP4: Small food portion and slow service", missing the "great taste" opinion on the "food" aspect of the comment. On the other hand, ABKPA, leveraging fine-grained ABSA to perform KPA at the aspect level, can identify and match every opinion expressed on the "food" and "service" aspects of the comment to single-aspect KPs, "KP1", "KP4" and "KP5" correctly, as shown in Table 1c. Neverthless, with both comments and KPs containing opinions on multiple aspects, sentence-based KPA also becomes ineffective and results in inaccrurate KP prevalence. For instance, in Table 1b, sentence-based KPA falsely map comment "1.1" and "3.1" with two overlapping KPs: "KP1" and "KP2", while both contain duplicate opinions on the same "service" aspect.

Our main contributions are: **(1)** We propose Aspect-based Key Point Analysis (ABKPA), a novel summarization framework for business reviews. ABKPA addresses the KPA shortcomings in sentence-based KP extraction and matching, which extract KPs with overlapping opinions and falsely matches KPs to long review comments containing multiple opinions. **(2)** Core to ABKPA is the use of fine-grained ABSA model to extract aspect-focused KPs without redundancy. **(3)** Importantly, using fine-grained ABSA tagging to automatically generate and annotate silver labels for aspect-sentiment matching examples, we employed contrastive learning and devised an aspect-based KP Matching model for more accurate KP quantification on business reviews.

2 Related Work

Based on the form of summaries, review summarization studies can be broadly grouped into three classes: Aspect-based Structured Summarization, Textual Summarization, and Key Point Analysis.

2.1 Aspect-based Structured Summarization

Early studies in the Data Mining community applied aspect-based sentiment analysis (ABSA) to extract, aggregate, and quantify opinions in reviews in the form of noun phrases (e.g., food, price, service) and positive and negative sentiment of the reviewed entity (Hu and Liu, 2004; Ding et al., 2008; Popescu and Etzioni, 2007; Blair-Goldensohn et al., 2008; Titov and McDonald, 2008). While these studies give basic quantification for reviews in terms of aspects and their sentiment, they lack textual explanation for the opinion details.

2.2 Textual Summarization

Document summarization is an important topic in the Natural Language Processing community, aiming to produce concise textual summaries capturing the salient information in source documents. While extractive review summarization approaches use surface features to rank and extract salient opinions for summarization (Mihalcea and Tarau, 2004; Angelidis and Lapata, 2018; Zhao and Chaturvedi, 2020), abstractive techniques use sequence-to-sequence models (Chu and Liu, 2019; Suhara et al., 2020; Bražinskas et al., 2020b,a; Zhang et al., 2020) to generate review-like summaries containing only the most prevalent opinions. Recently, prompted opinion summarization leveraging Large Language Models (LLMs) was applied to generate fluent and concise review summaries (Bhaskar et al., 2023; Adams et al., 2023). Still none of the existing studies focus on presenting and quantifying the diverse opinions in reviews.

2.3 Key Point Analysis

Originally developed to summarize arguments (Bar-Haim et al., 2020a,b), KPA was later applied to summarize and quantify the prevalence of opinions in reviews (Bar-Haim et al., 2021). Existing work on KPA for reviews has two major shortcomings. First, extraction of KPs relies on supervised models to identify short sentences with high argument quality as KPs, and such sentence-based extraction makes KPs often contain multiple and redundant opinions. Secondly, due to supervised training for the comment-KP matching model, despite containing multiple opinions, each comment is often mistakenly matched to a KP, leading to inaccurate quantification for KPs.

More recent research aims to generate high-level abstractive summaries for KPA. One class of studies (Cattan et al., 2023) is focused on structuring the KPs from extractive KPA as a hierarchy. Another class of studies is focused on abstractive summarization for KP generation (Kapadnis et al., 2021; Li et al., 2023b); an abstractive summarization model is employed to generate KPs either from each argument (Kapadnis et al., 2021), or by summarizing a cluster of arguments grouped by common theme (Li et al., 2023b). None of the recent studies focus on the core issues of KP redundancy KPs and inaccurate quantification for KPs.

3 Aspect-based Key Point Analysis

We propose the ABKPA framework, with the training and inference phases presented in Figure 1. ABKPA mainly leverages ABSA resources during its inference phase to enhance the quality of KPs through Aspect-based KP Extraction (Section 3.1), and precisely map comments with multiple opinions to various KPs via aspect-based KP Matching (Section 3.2). Notably, in the training phase, ABKPA again utilizes ABSA for automatic construction and labelling of aspect-sentiment matching pairs without human annotation (Section 3.3), which can effectively bootstrap our aspect-based KP Matching model through contrastive learning.

3.1 Aspect-based KP Extraction

Unlike argument summarization, short and good-quality comments are more frequent in business reviews, and they can be selected as KPs. Previous works use an argument quality ranking model to score and select KP candidates (Bar-Haim et al., 2020b). But it is not accurate for reviews because the quality was established to determine the magnitude of whether an argument supports/contests a controversial topic. Bar-Haim et al. (2021) then proposed a additional classifer to improve KP quality for review summarization, but the solution requires extra human annotation and computational resource. Also, using several ranking models lack generalizability because it is complex to hyper-tune the optimal thresholds for good KP selection. We filled this gap by defining aspect-based KP Extraction, which efficiently uses ABSA resources to eliminate short and highly-overlapping sentences in reviews and provide higher KP quality. Moreover, short sentences in reviews can also cover opinions on multiple aspects, whereas KPs with duplicate

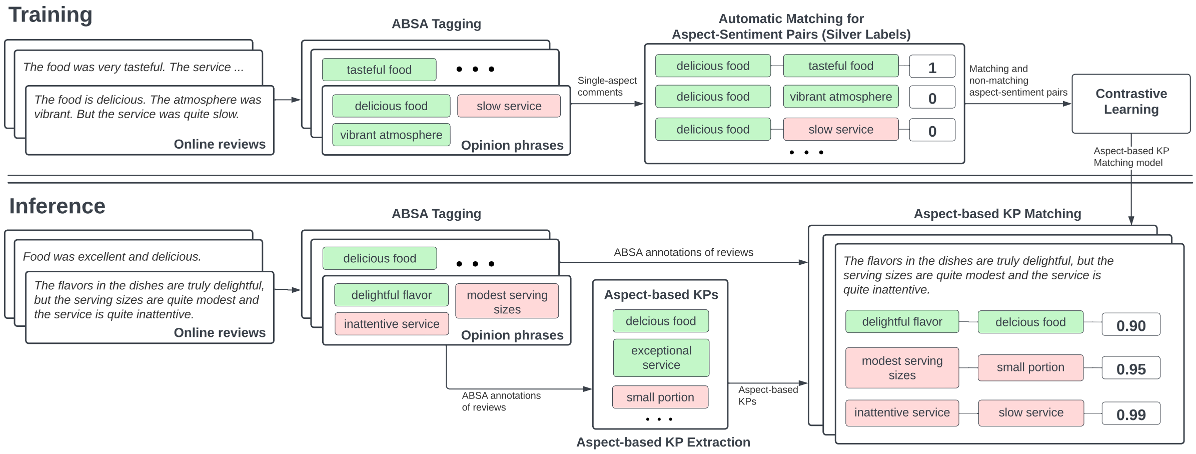
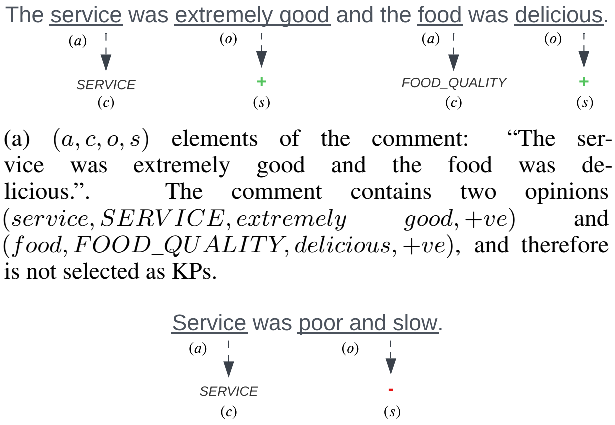

Training
The food was very tasteful. The service ...
The food is delicious. The atmosphere was vibrant. But the service was quite slow.
Online reviews
ABSA Tagging
tasteful food
delicious food
slow service
vibrant atmosphere
Opinion phrases
Single-aspect comments
Automatic Matching for Aspect-Sentiment Pairs (Silver Labels)
delicious food
tasteful food
1
delicious food
vibrant atmosphere
0
delicious food
slow service
0
Matching and non-matching aspect-sentiment pairs
Contrastive Learning
Aspect-based KP Matching model
Inference
Food was excellent and delicious.
The flavors in the dishes are truly delightful, but the serving sizes are quite modest and the service is quite inattentive.
Online reviews
ABSA Tagging
delicious food
delightful flavor
modest serving sizes
inattentive service
Opinion phrases
ABSA annotations of reviews
ABSA annotations of reviews
Aspect-based KPs
delcious food
exceptional service
small portion
Aspect-based KP Extraction
Aspect-based KPs
Aspect-based KP Matching
The flavors in the dishes are truly delightful, but the serving sizes are quite modest and the service is quite inattentive.
delightful flavor
delcious food
0.90
modest serving sizes
small portion
0.95
inattentive service
slow service
0.99

Figure 1: The training and inference phases of the ABKPA framework

opinions and aspects will affect the quantitative correctness of KP Matching. We address such limitations using aspect-based KP Extraction, which efficiently leverage ABSA resources to eliminate overlapping short sentences during KP Extraction and provide higher KP quality.

Existing studies developed fine-grained ABSA under different forms of elements (Pontiki et al., 2016; Wan et al., 2020). In this aspect-based KP Extraction task, we leverage elements from the (a, c, o, s) quadruple prediction of ABSA (Zhang et al., 2021), namely (a)spect term, aspect (c)ategory, (o)pinion term and (s)entiment polarity (positive or negative), to advance KP Extraction and KP Matching tasks in KPA.


The service was extremely good and the food was delicious.
(a)
(o)
(a)
(o)
SERVICE
(c)
+
(s)
FOOD_QUALITY
(c)
+
(s)
Service was poor and slow.
(a)
(o)
SERVICE
(c)
-
(s)

(a) (a, c, o, s) elements of the comment: “The service was extremely good and the food was delicious.”. The comment contains two opinions $(service, SERVICE, extremely\ good, +ve)$ and $(food, FOOD\_QUALITY, delicious, +ve)$, and therefore is not selected as KPs.

(b) (a, c, o, s) elements of the comment: “Service was poor and slow.”. The comment contains only one opinion $(service, SERVICE, poor\ and\ slow, -ve)$, and therefore is selected as KPs.

Figure 2: Elements of the quadruple prediction (a, c, o, s) of ABSA for two example comments (a) and (b), taken from Table 1. The examples also illustrate valid and invalid cases of KPs for reviews.

From examples in Figure 2, (a) is the aspect of a reviewed entity (e.g., *food*, *service*) on which users express their opinion (o), while (c) generalizes (a) into categories (e.g., *FOOD_QUALITY*), and (s) implies the attitude of (o) (e.g., *+ve*, or *-ve*).

We start by collecting high-quality KPs using the argument quality ranking model from (Bar-Haim et al., 2021), before performing ABSA prediction to retrieve the opinion phrases of all KP candidates. Then, we select only KPs having a single aspect and opinion, and sort KPs by descending order of their quality. Finally, we traverse the candidates from the list, target those sharing semantically similar opinion phrases and sentiments, and remove those with higher length yet lower quality from the list.

3.2 Aspect-based KP Matching Using Contrastive Learning

We devise an aspect-based KP Matching model in ABKPA, which directly scores the similarity of a single opinion of a comment towards extracted KP candidates. Our model is more effective than the traditional KP Matching model of sentence-based KPA because it can **(1)** bypass noise and redundancy in the full text, **(2)** capture and encode opinion information in long comments efficiently without having to truncate , and **(3)** better coordinate to the content of different aspects presented in the original comment. , based on extracted opinion phrases and sentiments. From Figure 3, aspect-based KP Matching employs contrastive learning to transform the original semantic embedding of a comment or KP into a new space where the position of positive matching pairs - with signals indicated by the (a, o, s) triplet of an opinion in comments

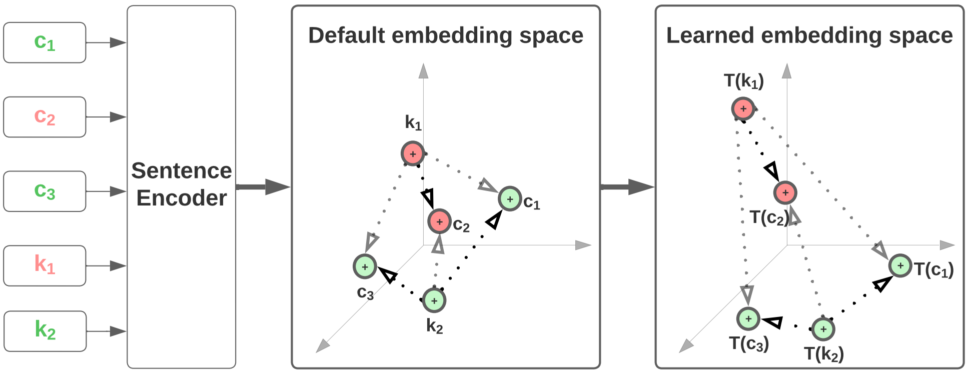


Figure 3: An example of the embedding space transformation. In this example, each node represents the opinion on a particular aspect of a comment (c) or key point (k), and is colored by their sentiments. The positive pairs (e.g., k_1 and c_2), whose (a, o, s) triplet of the opinions share a great similarity, are pulled closer to each other while negative pairs are pushed apart.

and KPs - are closer than negative pairs, and vice versa.

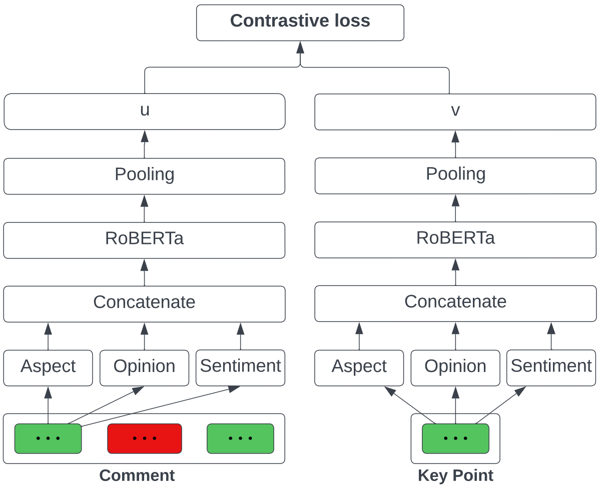


Figure 4: The siamese network architecture for training the comment-KP matching model of ASKPA

Figure 4 shows the siamese neutral network architecture for training the aspect-based KP matching model of ASKPA. We utilize the siamese neural network architecture, which was proven efficient for encoding of sentences (Reimers and Gurevych, 2019), for training the aspect-based KP Matching model. Formally, considering a single opinion from a comment (c) and key point (k), we create the training input as $\{T(\mathrm{c}), T(\mathrm{k}), label\}$, where $T(\mathrm{c})$ or $T(\mathrm{k})$ uses a special token <SEP> to concatenate tokens of the (a, o, s) triplet of an opinion from c or k, and $label$ is the matching silver label (0 or 1). For example:

$c =$ The staff is always courteous to customers

$T(c) =$ always courteous staff <SEP> positive

We then used a pre-trained language model to encode tokens in $T(\mathrm{c})$ and $T(\mathrm{k})$ of the pair. Then, we pass their embeddings through a siamese neural network, which is a mean-pooling layer to aggregate the token embeddings of each input into sentence embeddings. We compute the contrastive loss of sentence embeddings of each training input as:

$$\mathcal{L} = -y \cdot \log(\hat{y}) + (1 - y) \cdot \log(1 - \hat{y}) \quad (1)$$

where $\hat{y}$ is the cosine similarity of the embeddings, and y reflects whether a pair matches (1) or not (0). Using contrastive loss (Equation 1), the network is trained to encode the input sequences to make positive and negative examples more distinguishable in the new embedding space. During inference, sequences of single opinions from the comment-KP pairs are input into the network, and the cosine similarity is used to compute their matching score.

Because our new aspect-based KP Matching model utilizes the aspect-sentiment information, it also allows matching a comment with opinions on multiple aspects to various key points, which is more accurate than matching at the sentence (comment) level in sentence-based KPA (Bar-Haim et al., 2020b, 2021). During inference, given a comment and a set of aspect-based KPs, we first calculate the matching scores of opinions inside comments with all KP candidates, and then map every opinion to its best-matching KP.

To achieve effective contrastive learning for the aspect-based KP Matching model, comment-KP pairs annotated with positive (matching) and negative (non-matching) labels are needed. We present our approach to leveraging ABSA annotations to construct such training examples in Section 3.3.

3.3 Silver Label Annotation for KP Matching

Previous work relied on data from the argument domain to fine-tune the KP Matching model and apply cross-domain to business reviews. In this work, we sidestep the needs of crowdsourcing the training data for our aspect-based KP Matching model. Instead, ASKPA makes use of available ABSA resources from reviews to construct and annotate the training data for its aspect-based KP Matching model in the training phase. We formulate an annotation heuristic that autonomously produces and annotates matching pairs of comments and KPs into positive (matching) or negative (non-matching) labels. Such labels, terms "silver labels", derived from aspect-sentiment elements of comments/KPs, are crucial for training our aspect-based KP Matching model (Section 3.2)

Algorithm 1 Silver Label Annotation
Input: Comment c, KP Candidates K, Threshold t
Output: Generated positive and negative comment-KP pairs of c and key point in K

```
procedure ANNOTATE_SILVER_LABEL($s$, $K_{ac}$, $t$)
    $positive\_pairs \leftarrow []$
    $negative\_pairs \leftarrow []$
    for $k$ in $K$ do
        $asp\_c, opin\_c, pol\_c \leftarrow Get\_ABSA(c)$
        $asp\_k, opin\_k, pol\_k \leftarrow Get\_ABSA(k)$
        $cos\_asp\_c\_k \leftarrow Cos(asp\_c, asp\_k)$
        $cos\_asp\_k\_c \leftarrow Cos(asp\_k, asp\_c)$
        $cos\_asp \leftarrow Avg(cos\_asp\_c\_k, cos\_asp\_k\_c)$
        if $cos > t$ and $pol_c = pol_k$ then
            add $(c, k)$ to $positive\_pairs$
        else
            add $(s, k)$ to $negative\_pairs$
        end if
    end for
    return $positive\_pairs \cup negative\_pairs$
end procedure
```

Algorithm 1 presents the pseudo-code for generating and annotating silver labels for matching pairs in training samples. Firstly, note that in these training samples, we only include comments/KPs expressing their opinion on a single aspect. When provided with a comment and a set of aspect-based KPs extracted from a dataset D of a business category, e.g, hotels, restaurants, the algorithm annotates the matching labels from opinions of possible comment-KP pairs based on their (a)spect term, aspect (c)ategory, and (s)entiment (i.e., the (a, c, s) triplet). Formally, we give positive labels on constructed comment-KP pairs with:

$$\forall(\mathrm{c}, \mathrm{k}) \in \{c_i\}_{i=1}^{|D|},\ cos(\mathbf{e}^{a(\mathrm{c})}, \mathbf{e}^{a(\mathrm{k})}) \geq \theta, s(\mathrm{c}) = s(\mathrm{k})$$

where c and k are the comment and KP of the pair, $\mathbf{e}^{a(\mathrm{c})}$ and $\mathbf{e}^{a(\mathrm{k})}$ are the word embeddings of aspect terms from c and k, $\mathbf{s}(\mathrm{c})$ and $\mathbf{s}(\mathrm{k})$ are the sentiments from c and k, respectively, and $\theta \in (0, 1]$ is a threshold for deciding the homogeneity of the pair's aspect terms. We compute the cosine similarity of a pair's aspect terms as:

$$\cos(\mathbf{e}^{a(\mathrm{c})}, \mathbf{e}^{a(\mathrm{k})}) = \frac{\mathbf{e}^{a(\mathrm{c})^T}\mathbf{e}^{a(\mathrm{k})}}{||\mathbf{e}^{a(\mathrm{c})}||_2\ ||\mathbf{e}^{a(\mathrm{k})}||_2} \qquad (2)$$

We label the remaining pairs disqualified by the above matching criteria as negative pairs whose opinions have dissimilar aspects and/or sentiments.

4 Experiments

4.1 Experiment Setup

This experiment was designed to specifically assess the novel matching and modelling process of ABKPA over existing KPA studies. We compared the matching performance of ABKPA against the following SOTA KP Matching models:

RKPA: The sentence-based KP Matching model from the latest KPA study adapted for business reviews (Bar-Haim et al., 2021), which was trained using ArgKP - a KP Matching dataset on argument (Bar-Haim et al., 2020a).

RKPA+: An enhanced version for RKPA (Bar-Haim et al., 2021), where RKPA is fine-tuned using our aspect-sentiment matching examples with silver labels for training. We use this baseline to evaluate the effectiveness of silver-annotated training examples.

SMatch: A model using SMatchToPR - 1st ranked sentence-based KP Matching model for argument domain from the KPA-2021 shared task (Friedman et al., 2021). However, in this experiment, we fine-tuned it using our aspect-sentiment matching examples with silver labels for training. SMatch employs contrastive learning and sentence embedding for KP Matching but unlike ABKPA, it does not utilize aspect-sentiment information to measure the cosine similarity of comment-KP pairs. We use SMatch to evaluate the effectiveness of contrastive learning and also the efficiency of ABKPA over SMatch while aspect-sentiment information of comments and KPs is utilized for KP Matching.

Note that conventionally, RKPA, RKPA+, and SMatch can only match a comment to one best-matching KP, which makes them always fail to associate multiple KPs to comments with multiple opinions. In our experiment, for a fair comparison, we adjust these models to match every comment with top n highest-scored KPs, corresponding to the n opinion aspects identified in the comment.

ABKPA, together with the baseline models, were all fine-tuned on a RoBERTa-large model (Liu et al., 2019), using the Huggingface transformers framework. For hyperparameters, we used the optimal setting preferred by previous studies for the best results. We first pretrained all models with the Masked LM (MLM) task (Liu et al., 2019) to adapt it to reviews. The pretraining was performed for 2 epochs, a learning rate of 1e-5, following the procedure described by Bar-Haim et al. (2021). For ABKPA and SMatch, based on the setting of Alshomary et al. (2021), we fine-tuned the siamese network of the model for 10 epochs, with a batch size of 16, and a maximum input length of 128, leaving all other parameters to their defaults. For RKPA and RKPA+, we fine-tuned the KP Matching model for 9 epochs, with a learning rate of 5e-6, as suggested by (Bar-Haim et al., 2021), keeping all other settings at their default values. We trained all models using an NVIDIA GeForce RTX 3080Ti GPU. We implement the pre-trained model Snippext (Miao et al., 2020) to obtain ABSA predictions on review comments. For silver-annotation of reviews for matching, we employ SpaCy (Honnibal et al., 2020) to compute the cosine similarity of the aspect terms of constructed matching pairs.

4.2 Data

Following the latest KPA work (Bar-Haim et al., 2021), we used the popular Yelp Open Dataset [2] for empirical evaluation and we extended experiments to five business categories: *Arts & Entertainment* (25k reviews), *Automotive* (41k reviews), *Beauty & Spas* (72k reviews), *Hotels* (8.6K reviews), and *Restaurants* (680k reviews).

Each dataset, corresponding to a specific business category, was divided into 'training' and 'test' subsets. Reviews from the first and second top 30 most-commented business entities were sampled for training and evaluation, respectively. For both training and test subsets, we extract aspect-based KP candidates, constrained to 3-6 tokens, first following Bar-Haim et al. (2021) to compute the quality score of comments using the argument quality model (Toledo et al., 2019), with the minimum quality score 0.42. Then we applied extensive filters, discussed in Section 3.1, to retrieve aspect-based KPs for review summarization. Training samples were then constructed, and annotated for silver labels (discussed in Section 3.3) based on the remaining comments and the extracted aspect-based KPs.

Table 2: Annotations for test data in five dataset (i.e, business categories): Arts (& Entertainment), Auto(motive), Beauty (& Spas), Hotels, Restaurants.

Dataset	**# pairs**	**# +ve pairs**	**# KPs**
Arts	1536	69	32
Auto	877	93	18
Beauty	1093	77	22
Hotels	1680	72	35
Restaurants	1613	108	33

In the test subsets, for annotating the matching ground truth in test data (for evaluation), we used the Amazon Mechanical Turk [3] (MTurk) as the main crowdsourcing platform, based on the guideline of Bar-Haim et al. (2020a) and Bar-Haim et al. (2021). To prepare gold-labelled KPs in the test set for evaluation, we relied on human to annotate/select KPs. For each test subset, we guide annotators to select non-redundant KPs, prioritizing those with high-quality scores and fulfilling 4 properties of KPs for reviews (Bar-Haim et al., 2021), including *validity*, *sentiment*, *informativeness*, and *single-aspect*. Similarly, to ensure consistent quality in the test subsets, we limit to comments of 6-11 tokens. For each token length in this range, we select the top 8 highest-quality comments, creating a total of 48 comments per category. To annotate matching KP-comment pairs, we select from 8 annotations only those by annotators having high agreement with others (minimum κ score of 0.05). Details of the annotation scheme and quality control to ensure high-quality annotation are in Appendix A.

Table 2 summarises the statistics of the test data and their annotations for all categories. Overall, the test data has 6799 labelled (comment, KP) pairs, of which 419 pairs are positive. Note also that because the annotation covers the labels for all possible pairs, there are no undecided pairs.

4.3 Results

We fine-tuned and evaluated all models on the respective train and test subsets of different datasets (i.e, business category), except RKPA, which was fine-tuned on ArgKP, following the implementation of Bar-Haim et al. (2021). Our evaluation was based on the Average Precision (AP) used in the

[2] https://www.yelp.com/dataset

[3] https://www.mturk.com/

Table 3: AP score of KP Matching models. The best result of each experiment is in bold.

Dataset	All comments				Multiple-opinion comments			
	ABKPA	**SMatch**	**comm-Match**	**RKPA**	**ABKPA**	**SMatch**	**comm-Match**	**RKPA**
Arts	**0.99**	0.98	0.94	0.79	**0.99**	0.88	0.83	0.90
Auto	**0.77**	0.75	0.43	0.54	**0.80**	0.70	0.42	0.71
Beauty	**0.98**	0.97	0.84	0.62	**0.94**	0.88	0.81	0.62
Hotels	**0.99**	0.98	0.98	0.81	**0.93**	0.89	0.93	0.81
Restaurants	**0.87**	0.85	0.73	0.50	**0.83**	0.75	0.73	0.56
Average	**0.92**	0.91	0.78	0.65	**0.90**	0.82	0.74	0.72

Table 4: Model generalizability evaluation results. AP score in *out-of-category* experiment of KP Matching models, where data for one category is used for testing and models are trained on data for the rest categories. Note that no results for RKPA as it is trained on non-Yelp review data. The best result of each experiment is in bold. Result difference from the within-category experiment (Table 3) is shown in brackets, while (—-) indicates nil difference.

Dataset	All comments			Multiple-opinion comments		
	ABKPA	**SMatch**	**RKPA+**	**ABKPA**	**SMatch**	**RKPA+**
Arts	**0.98** (-.01)	0.95 (-.03)	0.90 (-.04)	**0.99** (—-)	0.80 (-.08)	0.83 (—-)
Auto	**0.76** (-.01)	0.51 (-.24)	0.40 (-.03)	**0.64** (-.12)	**0.64** (-.08)	0.41 (-.01)
Beauty	**0.94** (-.04)	0.97 (—-)	0.60 (-.24)	0.77 (-.17)	**0.84** (-.04)	0.54 (-.27)
Hotels	**0.98** (-.01)	0.96 (-.02)	0.92 (-.06)	**0.92** (-.01)	0.81 (-.07)	0.89 (-.04)
Restaurants	**0.87** (—-)	0.84 (-.01)	0.66 (-.07)	**0.75** (-.08)	0.61 (-.14)	0.69 (-.04)
Average	**0.91** (-.01)	0.85 (-.06)	0.70 (-.09)	**0.81** (-.08)	0.74 (-.08)	0.67 (-.04)

KPA-2021 shared task (Friedman et al., 2021) [4]. First, for all models, we extract the top 50% predicted matching pairs for each dataset by the order of their confidence (matching) score. Then, given the ground truth data, Average Precision (Turpin and Scholer, 2006) (AP), is calculated per dataset to evaluate the model matching performance. During evaluation, models are tested on two data configurations: "all comments" and "multiple-opinion comments", which explicitly aim to test the model's ability to handle comments with multiple opinions.

Table 3 presents the AP score for all models under "all comments" or "multiple-opinion comments" configurations. Overall, ABKPA shows the best performance, significantly outpacing other models (paired t-test, $p << 0.05$), with an average AP score of 0.92 and 0.90. Conversely, RKPA shows the lowest performance in three out of five datasets, mainly because it was fine-tuned with argument data and applied to reviews. RKPA+, sharing RKPA architecture but was fine-tuned using our silver-annotated reviews, display a higher performance overall. Finally, SMatch and ABKPA, by applying contrastive learning for KP Matching on the natural content of comments or on the opinion information of comments, respectively, achieve consistent improvements on all datasets. While both alternatives perform well and apply contrastive learning, ABKPA achieves higher and more consistent performance. This again demonstrates the benefit of integrating ABSA resources into ABKPA's KP Matching task.

In the "multiple-opinion comment" scenario, most models saw a certain performance decrease, mainly due to the long comments of multiple opinions challenging KP Matching. Surprisingly, RKPA shows a slight performance boost, likely benefiting from its extensive training data with longer sentences from the argument domain compared to our silver-annotated data. However, ABKPA still maintains its leading position with minimal performance variation.

4.4 Out-of-category experiment

In this set of experiments, we assess the generalizability of ABKPA and baseline models via out-of-category performance evaluation. Specifically, we test each model's performance on a dataset with a business category c (e.g., hotels), considering it was trained on all other datasets excluding c.

[4] https://2021.argmining.org/shared_task_ibm

Table 5: AP score of ABKPA and ABKPA$_{\neg C}$ on two test data settings.

Dataset	All comments		Multi-opinion comments	
	ASK-PA	**ASK-PA$_{\neg C}$**	**ASK-PA**	**ASK-PA$_{\neg C}$**
Arts	0.99	0.92	0.99	0.89
Auto	0.77	0.58	0.80	0.43
Beauty	0.98	0.85	0.94	0.82
Hotels	0.99	0.95	0.93	0.88
Restaurants	0.87	0.78	0.83	0.72

Table 4 presents the AP Score for all models in the out-of-category experiment. Comparing Table 3 and Table 4, the relative ranking of models remains similar, with ABKPA showing the best and most stable performance. In the "all comments" setting, ABKPA shows a very slight decrease in its AP Score (0.1 on average, drop varying from 0.01 to 0.04), while still outperforming other models significantly (paired t-test, $p < 0.05$), with an average AP score of 0.91. This shows that ABKPA can be generalised to new, unseen business categories. In contrast, SMatch and RKPA+ see notable performance drops – 0 to 0.24 for SMatch and 0.03 to 0.24 for RKPA+ – when transitioning from in-category to out-of-category, indicating their domain dependence, a finding aligned with existing studies. For multi-opinion comments, ABKPA remains the top performer with an AP score of 0.81 (compared to 0.74 for SMatch and 0.67 for RKPA+), while RKPA+ sees the most significant drop – from 0.04 to 0.27, emphasizing the instability of domain-dependent supervised training models.

4.5 Ablation study

Our ablation study examines the utility of contrastive learning for KP Matching. The ABKPA$_{\neg c}$ model, omitting constrastive learning, uses the positive and negative examples from our silver-annotated data to directly train a matching model. Table 5 highlights the performance disparity between ABKPA$_{\neg c}$ and ABKPA. Without contrastive learning, ABKPA$_{\neg c}$ exhibits a significant performance decline, highlighting the efficacy of contrastive learning in ABKPA. In the "all comments" setting, the average absolute AP score decreases by 0.10, ranging from 0.04 to 0.19. For "multi-opinion comments", the performance drop of ABKPA$_{\neg c}$ is even more pronounced, with the AP score declining from 0.90 to 0.75, varying from 0.05 to 0.37. These results demonstrate the importance of contrastive learning for the superb performance of ABKPA.

4.6 Case studies

We conduct a case study to evaluate KP redundancy on the "Restaurants" dataset, as shown in Table 7 (Appendix D). Overall, all baselines encounter redundancy (i.e., KPs with overlapping aspects and opinions) in the output. For example, the two KPs "The service here was exeptional." and "Customer service is excellent." contain redundant opinions because the baseline models lack the confidence to distinguish the better one while matching to comments. In contrast, ABKPA offers KP Matching with more diverse yet non-repetitive aspects.

We conduct another case study to evaluate the correctness of KP prevalence (i.e., salience score) of different models on popular KPs (i.e., KPs with a high number of comments in the ground truth). Table 8 (Appendix E) presents the prevalence computed by each model on the top three most prevalent KPs from each dataset. Note that in this case study, we only report the KP prevalence (i.e., salient score) computed in quantity by different models against actual prevalence, while ABKPA still has better matching performance than other baselines, as proved in Section 4.3. Overall, ABKPA achieves highly accurate KP prevalence and matching comments while being evaluated with the ground truth.

5 Conclusions

This paper proposed Aspect-based Key Point Analysis (ABKPA), a framework that effectively makes use of ABSA resources in business reviews to enhance multiple tasks of KPA. ABKPA addresses the major shortcoming of previous sentence-based KPA studies on the insufficient capture of comment's opinion and generation of redundant KPs. First, we leverage fine-grained ABSA to extract KPs by their aspects from comments, which significantly eliminates overlapping KPs compared to previous KPA studies. Secondly, leveraging ABSA for contrastive learning, we develop an effective aspect-based KP Matching model for mapping various KPs to comments with multiple opinions, which results in more accurate opinion quantification.

Limitations

The KP Matching model of ABKPA and other baselines was implemented using a RoBERTa large language model. Due to the high number of parameters (355M), the model requires high GPU resources for pre-training and fine-tuning. With limited GPU resource, we restrict the maximum input length of the baseline models to be 512 tokens. Moreover, the development, utilization of language model and reported performance assume the framework to be suitably implemented for English only.

Ethics Statement

We have applied ethical research standards in our organization for data collection and processing throughout our work.

The Yelp dataset used in our experiments was officially released by Yelp, which was published by following their ethical standard, after removing all personal information. The summaries do not contain contents that are harmful to readers.

We ensured fair compensation for crowd annotators on Amazon Mechanical Turk. We setup and conducted fair payment to workers on their annotation tasks/assignments according to our organization's standards, with an estimation of the difficulty and expected time required per task based on our own experience. Especially, we also made bonus rewards to annotators who exerted high-quality annotations in their assignments.

Acknowledgements

This research is supported in part by the Australian Research Council Discovery Project **DP200101441**.

References

Griffin Adams, Alex Fabbri, Faisal Ladhak, Eric Lehman, and Noémie Elhadad. 2023. From sparse to dense: GPT-4 summarization with chain of density prompting. In *Proceedings of the 4th New Frontiers in Summarization Workshop*, pages 68–74, Singapore. Association for Computational Linguistics.

Milad Alshomary, Timon Gurcke, Shahbaz Syed, Philipp Heinisch, Maximilian Spliethöver, Philipp Cimiano, Martin Potthast, and Henning Wachsmuth. 2021. Key point analysis via contrastive learning and extractive argument summarization. In *Proceedings of the 8th Workshop on Argument Mining*, pages 184–189, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Stefanos Angelidis and Mirella Lapata. 2018. Summarizing opinions: Aspect extraction meets sentiment prediction and they are both weakly supervised. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3675–3686, Brussels, Belgium. Association for Computational Linguistics.

Roy Bar-Haim, Lilach Eden, Roni Friedman, Yoav Kantor, Dan Lahav, and Noam Slonim. 2020a. From arguments to key points: Towards automatic argument summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4029–4039, Online. Association for Computational Linguistics.

Roy Bar-Haim, Lilach Eden, Yoav Kantor, Roni Friedman, and Noam Slonim. 2021. Every bite is an experience: Key Point Analysis of business reviews. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3376–3386, Online. Association for Computational Linguistics.

Roy Bar-Haim, Yoav Kantor, Lilach Eden, Roni Friedman, Dan Lahav, and Noam Slonim. 2020b. Quantitative argument summarization and beyond: Cross-domain key point analysis. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 39–49, Online. Association for Computational Linguistics.

Adithya Bhaskar, Alex Fabbri, and Greg Durrett. 2023. Prompted opinion summarization with gpt-3.5. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9282–9300.

Sasha Blair-Goldensohn, Kerry Hannan, Ryan McDonald, Tyler Neylon, George Reis, and Jeff Reynar. 2008. Building a sentiment summarizer for local service reviews.

Arthur Bražinskas, Mirella Lapata, and Ivan Titov. 2020a. Few-shot learning for opinion summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4119–4135, Online. Association for Computational Linguistics.

Arthur Bražinskas, Mirella Lapata, and Ivan Titov. 2020b. Unsupervised opinion summarization as copycat-review generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5151–5169, Online. Association for Computational Linguistics.

Arie Cattan, Lilach Eden, Yoav Kantor, and Roy Bar-Haim. 2023. From key points to key point hierarchy: Structured and expressive opinion summarization. *arXiv preprint arXiv:2306.03853*.

Eric Chu and Peter Liu. 2019. Meansum: A neural model for unsupervised multi-document abstractive summarization. In *International Conference on Machine Learning*, pages 1223–1232. PMLR.

Abhisek Dash, Anurag Shandilya, Arindam Biswas, Kripabandhu Ghosh, Saptarshi Ghosh, and Abhijnan Chakraborty. 2019. Summarizing user-generated textual content: Motivation and methods for fairness in algorithmic summaries. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–28.

Xiaowen Ding, Bing Liu, and Philip S Yu. 2008. A holistic lexicon-based approach to opinion mining. In *Proceedings of the 2008 international conference on web search and data mining*, pages 231–240.

Roni Friedman, Lena Dankin, Yufang Hou, Ranit Aharonov, Yoav Katz, and Noam Slonim. 2021. Overview of the 2021 key point analysis shared task. In *Proceedings of the 8th Workshop on Argument Mining*, pages 154–164, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177.

Manav Kapadnis, Sohan Patnaik, Siba Panigrahi, Varun Madhavan, and Abhilash Nandy. 2021. Team enigma at ArgMining-EMNLP 2021: Leveraging pre-trained language models for key point matching. In *Proceedings of the 8th Workshop on Argument Mining*, pages 200–205, Punta Cana, Dominican Republic. Association for Computational Linguistics.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Hao Li, Viktor Schlegel, Riza Batista-Navarro, and Goran Nenadic. 2023a. Do you hear the people sing? key point analysis via iterative clustering and abstractive summarisation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14064–14080, Toronto, Canada. Association for Computational Linguistics.

Hao Li, Viktor Schlegel, Riza Batista-Navarro, and Goran Nenadic. 2023b. Do you hear the people sing? key point analysis via iterative clustering and abstractive summarisation. *arXiv preprint arXiv:2305.16000*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Mary McGlohon, Natalie Glance, and Zach Reiter. 2010. Star quality: Aggregating reviews to rank products and merchants. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 4, pages 114–121.

Zhengjie Miao, Yuliang Li, Xiaolan Wang, and Wang-Chiew Tan. 2020. Snippext: Semi-supervised opinion mining with augmented data. In *Proceedings of The Web Conference 2020*, pages 617–628.

Rada Mihalcea and Paul Tarau. 2004. TextRank: Bringing order into text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 404–411, Barcelona, Spain. Association for Computational Linguistics.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad AL-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, Véronique Hoste, Marianna Apidianaki, Xavier Tannier, Natalia Loukachevitch, Evgeniy Kotelnikov, Nuria Bel, Salud María Jiménez-Zafra, and Gülşen Eryiğit. 2016. SemEval-2016 task 5: Aspect based sentiment analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 19–30, San Diego, California. Association for Computational Linguistics.

Ana-Maria Popescu and Orena Etzioni. 2007. Extracting product features and opinions from reviews. *Natural language processing and text mining*, pages 9–28.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Anurag Shandilya, Kripabandhu Ghosh, and Saptarshi Ghosh. 2018. Fairness of extractive text summarization. In *Companion Proceedings of the The Web Conference 2018*, pages 97–98.

Yoshihiko Suhara, Xiaolan Wang, Stefanos Angelidis, and Wang-Chiew Tan. 2020. OpinionDigest: A simple framework for opinion summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5789–5798, Online. Association for Computational Linguistics.

Wenyi Tay, Xiuzhen Zhang, and Sarvnaz Karimi. 2020. Beyond mean rating: Probabilistic aggregation of star ratings based on helpfulness. *Journal of the Association for Information Science and Technology*, 71(7):784–799.

Ivan Titov and Ryan McDonald. 2008. A joint model of text and aspect ratings for sentiment summarization. In *proceedings of ACL-08: HLT*, pages 308–316.

Assaf Toledo, Shai Gretz, Edo Cohen-Karlik, Roni Friedman, Elad Venezian, Dan Lahav, Michal Jacovi, Ranit Aharonov, and Noam Slonim. 2019. Automatic argument quality assessment - new datasets and methods. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5625–5635, Hong Kong, China. Association for Computational Linguistics.

Andrew Turpin and Falk Scholer. 2006. User performance versus precision measures for simple search tasks. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 11–18.

Hai Wan, Yufei Yang, Jianfeng Du, Yanan Liu, Kunxun Qi, and Jeff Z Pan. 2020. Target-aspect-sentiment joint detection for aspect-based sentiment analysis. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 9122–9129.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pages 11328–11339. PMLR.

Wenxuan Zhang, Yang Deng, Xin Li, Yifei Yuan, Lidong Bing, and Wai Lam. 2021. Aspect sentiment quad prediction as paraphrase generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9209–9219, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Chao Zhao and Snigdha Chaturvedi. 2020. Weakly-supervised opinion summarization by leveraging external information. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9644–9651.

A Annotation and Labelling Details of Test Data

For labelling the matching pairs on the test data for evaluation, we mainly annotate data using the Amazon Mechanical Turk [5] (MTurk) crowdsource platform, based on the guidelines of Bar-Haim et al. (2020a) and Bar-Haim et al. (2021). To ensure annotation quality, we only select workers with $\geq$ 80% lifetime approval rate and have at least 10 annotations approved. For each comment, annotators were prompted to select none or multiple relevant key points, where they are not exposed to any ABSA information to ensure fair evaluation of all models and not to favour ABKPA. Note also that each comment was labeled by 8 annotators, and they can freely decide the number of matching key points to a comment. Further, following Bar-Haim et al. (Bar-Haim et al., 2021), we ignore the judgement of annotators whose annotator-κ score < 0.05. This score averages all pair-wise Cohen's Kappa (Landis and Koch, 1977) for a given annotator, for any annotator sharing at least 50 judgments with at least 5 other annotators. Details of the annotation task description and guidelines for the crowd-workers are provided in Appendix B.

We consolidate the labels for every matching pair following Bar-Haim et al. (Bar-Haim et al., 2020a), where the *agreement score* for a comment-KP pair – the fraction of annotations as matching – is used to select positive and negative pairs. We decided to label comment-KP pair as **(i)** positive if the agreement score $> 30\%$, **(ii)** negative if agreement score $< 15\%$; and **(iii)** otherwise undecided. Note that there are no undecided pairs because the annotation covers the labels for all possible pairs. Note also that the agreement score threshold of 30% for labelling positive pairs is different from the 60% threshold used for argument data by Bar-Haim et al. (Bar-Haim et al., 2020a)) and is set empirically. Details of the experiment are provided in Appendix C.

B Key Point Matching Annotation Guideline of Test Data

We report details of the annotation task description and instruction to the Amazon Mechanical Turk crowd-workers as follows:

Task title: Match the review sentence to its relevant key point(s)

Task description: Workers are required to mark valid key point(s) (short, high-quality, and concise sentences) that represent the content of a sample sentence

Instruction:

In this task you are presented with a business domain, a sentence taken from a review of a business in that domain and a key point.

Choose multiple key points that represent the content (of mentioned aspects) in the given sentence.

Note that a sentence might cover opinions on multiple aspects of the reviewed entity. Please select all relevant KPs that represent all aspects mentioned in the sentence.

C Analysis of Agreement Score for Positive Label on Test Data Annotation

We use an agreement score threshold of 30% for labelling positive pairs for reviews, different than the 60% used for argument data by Bar-Haim et al.

[5] https://www.mturk.com/

Table 6: Percentage of comments by key point matches by different agreement score for matching pairs

Agreement score	No key point	Ambiguous	Single KP	Multiple KP
0.1	0.42%	0%	2.08%	97.50%
0.2	2.08%	0%	20.83%	77.08%
0.3	**5.83%**	**3.33%**	**40.00%**	**50.83%**
0.4	6.25%	13.75%	53.75%	26.25%
0.5	6.25%	13.75%	53.75%	26.25%
0.5	2.08%	35.42%	53.75%	8.75%

(2020a)). For business reviews, because sentences are shorter and are more likely to contain overlapping opinions than online argument debates, annotators tend to select more KPs to match a comment. For example, the annotators might match the comment *“waitress was very polite"* to either or both *“staff is courteous"*, and *“servers are great"* key points, and have less consistent annotations. Table 6 shows the percentage of comments by key point matches using different thresholds t for the agreement score within 0.1-0.6. In this measurement, a comment is matched to a key point if at least t annotators agree. Similarly, a comment has no key point if at least t annotators match it to 'None'. Otherwise, the comment is 'ambiguous'. From Table 6, we observe a tradeoff between the number of positive comment-KP pairs and the agreement score. As soon as the agreement score threshold is above 0.3, there are more comments with insufficient confidence in their annotations while matching with key points, resulting in a high proportion of ambiguous cases. We, therefore, use 0.3 as the threshold for the agreement score. Interestingly, from Table 6, key points selected by humans can cover about 90% of comments, with 50.83% of the comments mapped to more than one key point, showing the quality of our annotation for comments with multiple aspects.

D KP Summary Output

This section presents details of Table 7, which shows the top 5 negative KPs for all models, ranked by their prevalence, for the Hotels domain,

E KP Matching Prevalence Output

This section presents details of Table 8, which shows the performance of different models in our case study on the top three important KPs in every dataset.

Table 7: Top 6 positive-sentiment key points ranked by their predicted prevalence on "Restaurants" datasets. While ABKPA generates distinct KPs on single aspects, baseline models generate KPs with overlapping aspects and opinions. KPs that overlap with higher-ranked ones (i.e., KPs with higher prevalence) are noted with a (*redundant*) postfix

ABKPA	**SMatch**	**RKPA+**	**RKPA**	**ABKPA$_{\neg C}$**
Staff was courteous and accommodating.	Staff was courteous and accommodating.	Staff was courteous and accommodating.	Employees are friendly and attentive.	Staff was courteous and accommodating.
Generous sized portions.	Prices are fair and reasonable.	The service here was exceptional.	The service here was exceptional.	Fresh food , using local produce.
Service was prompt and friendly.	Fresh food , using local produce.	Fresh food , using local produce.	Ambiance is casual and comfortable.	Customer service is excellent.
Fantastic drink selection.	The service here was exceptional.	The food is consistently excellent!	Fresh food , using local produce.	The service here was exceptional. (*redundant*)
Prices are fair and reasonable.	Generous sized portions.	Customer service is excellent. (*redundant*)	Really delicious food , well balanced!	Lots of outdoor seating.
Delicious and expertly prepared food.	Service was prompt and friendly. (*redundant*)	Prices are fair and reasonable.	Staff was courteous and accommodating. (*redundant*)	Amazing authentic flavor!

Table 8: Prevalence on important key points (top three most common KPs among the framework) comparing with the ground truth.

#	Key Point	ABKPA	SMatch	comm-Match	RKPA	AS-KPA$_{\neg c}$	Human
Arts (& Entertainment)							
1	Friendly and helpful staff.	10	10	12	10	10	14
2	Seats are adequately comfortable.	4	6	4	5	4	4
3	Horrible customer service.	2	3	2	3	3	3
Auto(motive)							
1	They have excellent customer service.	6	7	1	4	10	29
2	The employees here are wonderful!	3	2	1	12	2	13
3	Very professional staff	4	5	3	2	0	13
Beauty (& Spas)							
1	Staff is friendly and accomodating.	14	14	33	6	13	18
2	Customer service- Excellent!	5	5	4	2	7	13
3	Amazing & professional service.	3	1	4	24	3	14
Hotels							
1	Friendly and helpful staff.	19	15	16	19	16	21
2	Clean and comfortable rooms.	9	10	8	11	12	13
3	The ambiance is wonderfully peaceful	1	2	3	0	2	1
Restaurants							
1	Staff was courteous and accomodating.	10	12	10	3	11	19
2	Fresh food, using local produce.	5	5	7	3	8	5
3	The service here was exceptional	2	5	6	6	5	5