FineSum: Target-Oriented, Fine-Grained Opinion Summarization

Suyu Ge
Department of Computer Science,
University of Illinois at Urbana-Champaign,
Urbana, IL, USA
suyuge2@illinois.edu

Jiaxin Huang
Department of Computer Science,
University of Illinois at Urbana-Champaign,
Urbana, IL, USA
jiaxinh3@illinois.edu

Yu Meng
Department of Computer Science,
University of Illinois at Urbana-Champaign,
Urbana, IL, USA
yumeng5@illinois.edu

Jiawei Han
Department of Computer Science,
University of Illinois at Urbana-Champaign,
Urbana, IL, USA
hanj@illinois.edu

ABSTRACT

Target-oriented opinion summarization is to profile a target by extracting user opinions from multiple related documents. Instead of simply mining opinion ratings on a target (e.g., a restaurant) or on multiple aspects (e.g., food, service) of a target, it is desirable to go deeper, to mine opinion on *fine-grained* sub-aspects (e.g., fish). However, it is expensive to obtain high-quality annotations at such fine-grained scale. This leads to our proposal of a new framework, *FineSum*, which advances the frontier of opinion analysis in three aspects: (1) *minimal supervision*, where no document-summary pairs are provided, only aspect names and a few aspect/sentiment keywords are available; (2) *fine-grained opinion analysis*, where sentiment analysis drills down to a specific subject or characteristic within each general aspect; and (3) *phrase-based summarization*, where short phrases are taken as basic units for summarization, and semantically coherent phrases are gathered to improve the consistency and comprehensiveness of summary. Given a large corpus with no annotation, FineSum first automatically identifies potential spans of opinion phrases, and further reduces the noise in identification results using aspect and sentiment classifiers. It then constructs multiple fine-grained opinion clusters under each aspect and sentiment. Each cluster expresses uniform opinions towards certain sub-aspects (e.g., "fish" in "food" aspect) or characteristics (e.g., "Mexican" in "food" aspect). To accomplish this, we train a spherical word embedding space to explicitly represent different aspects and sentiments. We then distill the knowledge from embedding to a contextualized phrase classifier, and perform clustering using the contextualized opinion-aware phrase embedding. Both automatic evaluations on the benchmark and quantitative human evaluation validate the effectiveness of our approach.[1]

[1]Our code and annotation are available at https://github.com/gesy17/FineSum.

WSDM '23, February 27-March 3, 2023, Singapore, Singapore

ACM ISBN 978-1-4503-9407-9/23/02…$15.00
https://doi.org/10.1145/3539597.3570397

Figure 1: An example of fine-grained opinion summarization of a restaurant (Opinion clusters are separated by dot lines).

KEYWORDS

Opinion Summarization, Aspect Extraction, Sentiment Analysis

1 INTRODUCTION

Target-oriented opinion summarization is to profile a target by aggregating user opinions on different aspects from multiple documents (e.g., profiling a restaurant from online reviews). It benefits intelligent decision making by succinctly displaying diverse opinions to users and reducing the information overload.

Different from generic multi-document summarization, the large volumes of reviews and the inherent subjectivity within them pose challenges to curating golden annotation for this task, rendering end-to-end training infeasible. A majority of work focuses on developing weakly-supervised or unsupervised summarization approaches [9, 35]. To further handle the diversity and conflicts in user opinions, some approaches perform aspect extraction and sentiment polarization at first, then generate sentence-level summaries for different aspects and sentiments in either extractive or abstractive forms [3, 18].

Though previous methods partly consider the heterogeneity in user opinions, we argue that they still summarize at a coarse level for two reasons: (1) Opinions in the same ⟨aspect, sentiment⟩ category may target at different subjects (e.g., in the ⟨food, good⟩ category of Fig. 1, one set of opinions can be about "burger" but the

other about "fish"), and (2) different opinions may focus on different characteristics of the same subject (e.g., one may praise waiters for their kind service but the other may comment on their slowness). By generating uniform summary, traditional methods overlook the diversity and conflicts in reviews within the same general aspect, which sometimes leads to ambiguous and inconsistent summary [2]. Motivated by this, we propose to drill down to sub-aspect level. For a specific aspect, we automatically discover subjects and their different attributes as sub-aspects and aggregate similar opinions for them. However, it's common to see multiple sub-aspects entangled in the same sentence, rendering traditional extractive and abstractive methods sub-optimal for this task. For example, the sentence "I was impressed by the warm-hearted waiters even though they are sometimes slow." In this scenario, extracting or excluding the whole sentence from a sub-aspect will either bring outliers or cause information loss; whereas abstractive summarization usually suffers from hallucinations [15, 24, 37] or even distortion of opinions. Therefore, instead of following traditional paradigm in text generation, we propose to summarize opinions with finer semantic units. We define a sequence of words containing an object and its description as an *opinion phrase*. Such phrases are leveraged as the basic components of our fine-grained summarization to reduce vagueness and hallucinations in extracted user opinions.

In this paper, we propose *FineSum*, a weakly-supervised framework for target-oriented, fine-grained opinion summarization. FineSum consists of the following three stages: **(1)** *Extracting candidate opinion phrases* by identifying objects and associated description from the raw corpus. **(2)** *Identifying possible aspect and sentiment* in each candidate phrase with two complementary models: (i) the opinion-oriented spherical embedding and (ii) the contextualized language model classifier fine-tuned for opinion identification. Since there is no available annotation, we first propose an opinion-oriented embedding to explicitly represent aspects and sentiments along with word semantics in a distinctive sphere space, which reflects the class probability of a phrase from its directional similarity with each aspect and sentiment embedding. Meanwhile, as the spherical embedding is context-free, we propose to transfer the learned distributional distinctiveness to a language model and evokes its capacities for contextualized phrase modeling. To exclude noisy no-class phrases that are inevitably introduced in stage (1), we additionally enhance the language model via robust model ensemble. **(3)** *Aggregating phrases within each aspect and sentiment* to obtain fine-grained opinion clusters. We represent phrases using the opinion-aware embeddings from the language model and cluster them by looking at the embedding similarity.

We summarize our contributions as follows:

- We propose the first systematic approach for target-oriented fine-grained opinion summarization, where opinion analysis for a target is comprehensive and drills down to sub-aspects. Specifically, we propose to solve the task with minimal supervision by leveraging only aspect names and a few keywords.
- To classify phrases without any training data, we propose to first distinguish different opinion semantics with an opinion-oriented spherical embedding space, then distill its knowledge to a language model for better context modeling, and finally integrate them for robust classification.

2 RELATED WORK

Previous work on unsupervised or weakly-supervised opinion summarization mainly adopts a popularity-based approach (i.e., extract or generate sentences containing the most salient opinions in the original corpus) [6, 11, 16, 17]. They usually select the most salient opinions to reduce redundancy in the generated summary. A majority of early methods focus on extractive summarization. Ku et al. [19] define popular opinions using TF-IDF and use pre-defined keyword sets to retrieve the most relevant and opinionated sentences. Paul et al. [30] extract opinions according to a variety of lexical and syntactic features, and calculate salience and contrastiveness of sentences using random walk. Recently, with the proliferation of end-to-end training, abstractive summarization receives much more attention. A typical practice is to encode salient information using an aggregated representation, then output new sentences by reconstructing this representation. A representative method is MeanSum [4], which generates summaries by training an auto-decoder to reconstruct the averaged input representation. Similarly, Amplayo and Lapata [1] condense review documents into multiple dense vectors and use a multi-source fusion module to generate summaries. Though effective in distilling the most salient information in reviews, popularity-based summarization methods suffer from information loss since they only consider popular opinions. Moreover, they overlook the heterogeneity and conflicts in opinions by generating uniform summaries for all types of opinions.

To overcome the weakness of popularity-based methods, a few methods propose to generate stance-aware summarization. Oved and Levy [29] guarantee the diversity of opinions by generating multiple different summaries per product via systematic perturbations. However, the generated abstractive summaries are not always controllable and interpretable due to lack of explicit aspect categorization. To enhance explainability, other work tries to summarize according to different aspects and sentiments [10, 18], which are more relevant to our method. They identify different aspects and sentiments using light supervision signals, such as product domain labels, user-provided ratings, and keyword sets. Angelidis and Lapata [3] first couple extractive opinion summarization with the tasks of aspect identification and sentiment polarization. It uses an aspect extractor trained under a multi-task objective and a sentiment predictor based on multiple instance learning. Following this, Angelidis et al. [2] represent aspects as discrete latent codes in the quantized transformer space, and encode sentences to the aspect space using a variational autoencoder. However, the above methods remain coarse-grained because they are not designed to explicitly capture sub-aspects and specific characteristics within each aspect. On the contrary, we propose to generate fine-grained phrase clusters for sub-aspects. To our knowledge, this is the first work attempting to generate fine-grained summarization inside each aspect and sentiment in the form of opinion phrase clusters.

3 PROBLEM DEFINITION

Given a corpus T containing reviews for targets $\{t_1, t_2, \ldots\}$ from a single domain (e.g., restaurants), we define a domain-related aspect set $A = \{a_1, a_2, \ldots\}$ and sentiment set $S = \{s_1, s_2, \ldots\}$ and input a keyword list as L_a or L_s for each a or s. For every target t, we

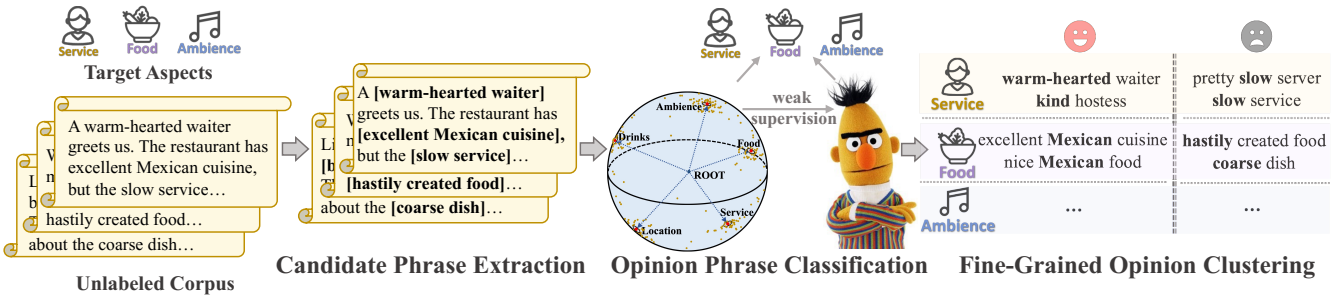


Figure 2: An overview of FineSum: Our proposed target-oriented, fine-grained opinion summarization framework.

define its review sentence set as $R = \{r_1, r_2, \ldots\}$, where each review r consists of multiple sentences $(x_1, x_2, \ldots)$. Each phrase p is defined as a non-overlapped word sequence $(w_1, w_2, \ldots)$ in one sentence. For each target, our final model outputs are a set of clusters $C = \{c_1, c_2, \ldots\}$ for every aspect-sentiment pair $\langle a, s \rangle$, where each cluster c contains multiple semantically coherent phrases $(p_1, p_2, \ldots)$.

4 FINESUM: OUR FRAMEWORK

Fig. 2 illustrates the overall workflow of the FineSum framework, which decomposes the task of aspect-based fine-grained opinion summarization into three stages. First, *candidate phrase extraction*, performs syntactic analysis to bring up multi-word sequences that potentially contain opinions as candidate phrases (Sec. 4.1). Second, *opinion phrase classification*, aims to classify extracted phrases into different aspects and sentiments (Sec. 4.2). The last stage, *opinion phrase mining*, generates fine-grained clusters within each aspect and sentiment by gathering semantically coherent phrases into the same cluster (Sec. 4.3). We introduce them as follows.

4.1 Candidate Phrase Extraction

Extracted Candidate Phrases

We went there [last night]. [No allergic reactions]. The [shrimp tacos and house fries are my standbys]. The [fries are sometimes good and sometimes great], and the [spicy dipping sauce they come with is to die for]. [Full beer menu] and [long cocktail lists], all [reasonable prices].

Figure 3: An illustration of extracted phrases. We highlight the subject and its associated descriptions inside each [phrase]. Best viewed in color.

As the first step, we seek to collect concise but complete semantic units in the original review sentences as the basic components of our opinion clusters. We observe that opinions in a sentence can usually be represented by one or more short phrases within it. For instance, the sentence "I went there last Sunday to order my favourite crispy chips and enjoy their heartful service." can be shorten to two phrases "favourite crispy chips" and "heartful service". Thus we propose to extract such "opinion phrase" as our summarization unit, which can be formally defined as below:

Definition of Opinion Phrase: A consecutive sequence of words in a sentence that includes subject(s) and users' feelings or descriptions of the subject(s).

However, existing phrase mining methods are usually designed to extract entity alike structures [8, 13, 33], which differs from our goal of extracting opinions. To mine opinions, we need to simultaneously discover a *noun subject*, along with its associated *adjective/adverb/verb descriptions*. To recognize such syntactic patterns, we designed a light-weight method using custom parsing tools. Specifically, we identify one-hop relation between a *noun* and *adjectives/adverbs* revealed by dependency parsing[2]. Besides, we capture co-existing *noun* component and *verb* component in the same level of the constituency parsing tree[3]. As our goal is to discover sufficient candidate phrases, we count the union of their results towards our final candidate phrase set, aiming for a phrase extractor with high recall. We show an example of phrase extraction results in Fig. 3.

4.2 Opinion Phrase Classification

The phrase extraction method described above retrieves a large number of candidate phrases according to their syntactic structure, but it overlooks the semantic meaning of phrases, resulting in a large but noisy candidate set. For example, the phrase "last night" in Fig. 3 is irrelevant to any aspect and sentiment. To exclude such unrelated phrases, and more importantly, identify the correct aspect and sentiment of each remaining phrase, this stage of FineSum aims to train accurate phrase-level aspect and sentiment classifiers. To achieve this without annotation, we first leverage the set of provided keywords as seeds to learn text embeddings tailored for opinions (i.e., aspects and sentiments). To better utilize context information, we further distill the learned opinion discriminative signals to pretrained language model through sentence-level fine-tuning. Then we ensemble the wisdom from embedding space and language model by adopting a self-training strategy on newly-encountered phrases. Each step of classification is introduced in detail below. For simplicity, we only take aspect classification as an example. The sentiment classification follows the same procedure.

4.2.1 Learning Opinion-oriented Embedding. Given the aspect set and keyword lists, we leverage them as seeds to learn text embeddings tailored for aspects through a generative process. This enables us to weigh the relevance of tokens with aspects directly from their

[2]https://stanfordnlp.github.io/CoreNLP/depparse.html
[3]https://stanfordnlp.github.io/CoreNLP/parse.html

embedding similarity. Meng et al. [26] shows that learning spherical distribution is more suitable to model topic-level similarity than regularizing Euclidean distribution due to the convenience to impose topic similarity constraints on spheres. The superiority of modeling topics as spherical distribution has also been proved in other tasks, including event clustering [34] and topic modeling [25]. Motivated by them, we design a spherical text embedding model for weakly-supervised aspect classification, where each aspect is surrounded by its representative keywords and sentences on the sphere. In this way, the semantic of a word or sentence is explicitly measured by its directional similarity with different aspects. Specifically, we design the corpus generation process according to the following three considerations: (1) *Inter-aspect Distinctiveness*: Different aspects a_i and a_j should locate far from each other in the embedding space; (2) *Intra-aspect Cohesiveness*: For each aspect a_i, its aspect-indicative words w_{a_i} and sentences x_{a_i} should distribute around a_i; and (3) *Context Dependency*: Words appearing within the same local context window and global context (sentence and aspect) should share similar representations.

Accordingly, the learning goal of our corpus generation process $\mathcal{L}_{corpus}$ is set as follows:

$$\begin{aligned} \forall a, x, w, \quad & ||\mathbf{a}|| = ||\mathbf{x}|| = ||\mathbf{w}|| = 1, \\ \mathcal{L}_{corpus} = & -\mathcal{L}_{inter} - \mathcal{L}_{intra} - \mathcal{L}_{context}, \\ \mathcal{L}_{inter} = & \sum_{a_i \in A} \sum_{a_j \in A \setminus \{a_i\}} min(0, 1 - \mathbf{a}_i^T \mathbf{a}_j - m_{inter}), \\ \mathcal{L}_{intra} = & \sum_{a_i \in A} \sum_{w_j \in L_{a_i}} min(0, \mathbf{w}_j^T \mathbf{a}_i - m_{intra}), \\ \mathcal{L}_{context} = & \sum_{x \in T} \log p(\mathbf{x}|\mathbf{a}_x) + \sum_{x \in T} \sum_{w_i \in x} \log p(\mathbf{w}_i|\mathbf{x}) + \\ & \sum_{x \in T} \sum_{w_i \in x} \sum_{|j| \le h, j \ne 0} \log p(\mathbf{w}_{i+j}|\mathbf{w}_i), \end{aligned} \tag{1}$$

where m_{inter} and m_{intra} are two learnable parameters, and h is the context window length. A is the given aspect sets and T is the corpus containing all sentences. L_{a_i} is the keyword list of the aspect a_i. To explain, the first objective $\mathcal{L}_{inter}$ encourages inter-aspect distinctiveness across different aspects by enforcing the cosine distance between any two aspects $\mathbf{a}_i$ and $\mathbf{a}_j$ to be larger than m_{inter}. The second objective $\mathcal{L}_{intra}$ requires the embeddings of aspect keywords $\mathbf{w}_j$ to be placed near the aspect center direction $\mathbf{a}_i$ within a local region m_{intra}. The third objective $\mathcal{L}_{context}$ models the context generation process conditioned on aspects in a three-step process: (1) $p(\mathbf{x}|\mathbf{a}_x)$ conditions each sentence $\mathbf{x}$ on a sampled aspect $\mathbf{a}_x$, (2) $p(\mathbf{w}_i|\mathbf{x})$ models the semantic coherence between a word $\mathbf{w}_i$ and the sentence $\mathbf{x}$ it appears, and (3) $p(\mathbf{w}_{i+j}|\mathbf{w}_i)$ models neighbor words $\mathbf{w}_{i+j}$ and $\mathbf{w}_i$ within local contexts.

As an aspect can be further divided into multiple fine-grained semantics, instead of modeling the aspect embedding $\mathbf{a}_x$ as a single vector, we assume a spherical von Mises-Fisher (vMF) distribution [20] for each aspect, and generate sentences and words by sampling from the distribution. Specifically, taking the generation of the aspect vector $\mathbf{a}_x$ as an example, it is parameterized by a mean vector $\mathbf{c}_i$ and a concentration parameter κ_{c_i}, which are specific to each aspect:

$$p(\mathbf{a}_x, \mathbf{c}_i, \kappa_{c_i}) = n_p(\kappa_{c_i}) \exp(\kappa_{c_i} \cdot \cos(\mathbf{a}_x, \mathbf{c}_i)), \tag{2}$$

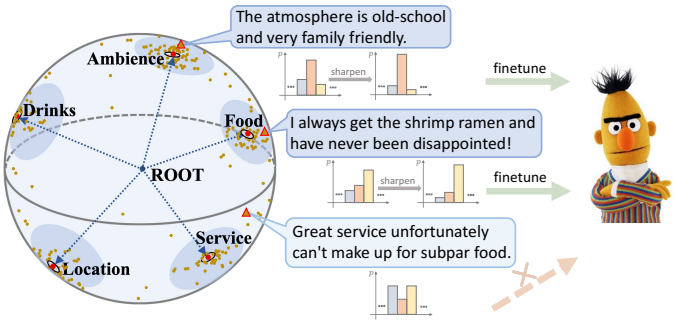


Figure 4: An illustration of knowledge distillation to contextualized classifier. Training sentences are selected from the dark circle area around each aspect. Best viewed in color.

where $n_p(\kappa_{c_i})$ is a normalization constant, $\cos(\mathbf{a}_x, \mathbf{c}_i)$ measures the cosine similarity between $\mathbf{a}_x$ and $\mathbf{c}_i$. Then the conditional probability $p(\mathbf{x}|\mathbf{a}_x)$ of a sentence $\mathbf{x}$ under its true aspect $\mathbf{a}_x$ becomes:

$$p(\mathbf{x}|\mathbf{a}_x) \propto \prod_{w_i \in x} p(\mathbf{w}_i|\mathbf{a}_x) \propto \prod_{w_i \in x} \exp(\cos(\mathbf{w}_i, \mathbf{a}_x)). \tag{3}$$

Similarly, the conditional probabilities $p(\mathbf{w}_i|\mathbf{x})$ and $p(\mathbf{w}_{i+j}|\mathbf{w}_i)$ can be computed by measuring their cosine similarities.

For model optimization, we develop a principled EM optimization procedure. Generally, the E-step estimates new aspect assignment of words with the updated parameters, and the M-step optimizes model parameters according to the learning goal in Eq. (1). Initially, only a few aspect-indicative keywords ($w_j \in L_{a_i}$) are provided as supervision for the M-step. The model then continuously refines the embedding space iteratively.

4.2.2 Distilling Knowledge to Contextualized Classifier. Context is crucial for phrase aspect classification. The context-free spherical embedding space mainly captures word-level discriminative signals but is insufficient to model sequential information from ordering of words. Therefore, we propose to distill knowledge from the aspect-regularized embedding space to a pre-trained language model, and leverage it as a contextualized phrase classifier. However, there are two challenges in this design: (1) The spherical embedding mainly captures the semantics of words and sentences, while it fails to directly reflect the meaning of phrases; (2) Word embedding inherently encodes different types of word-level semantics, which will introduce irrelevant information to aspect classification and probably result in an error-prone classifier. To solve both challenges, we propose to fine-tune the language model only on sentences with high aspect confidence. The underlying assumption is that a sentence classifier with high performance should also perform well on phrases. Specifically, we leverage soft predictions given by the directional similarity between sentence embeddings and aspect embeddings as weak supervision to fine-tune the BERT-base model [5] for aspect classification. The fine-tuning workflow is illustrated in Fig 4. To ensure the high quality of supervision signals, we only select the top-K nearest sentences around each aspect a_i, i.e., K sentences with the highest $\cos(\mathbf{x}, \mathbf{a}_i)$ scores. We "sharpen" the directional similarity to encourage better aspect separation, and transform it into pseudo training labels $\mathbf{l}_x$ as below:

$$l_{xi} = \frac{\exp(\alpha \cdot \mathbf{x}^T \mathbf{a}_i)}{\sum_{a_i \in A} \exp(\alpha \cdot \mathbf{x}^T \mathbf{a}_i)}, \tag{4}$$

where l_{xi} is the probability of sentence x belonging to the i_{th} aspect. α is the temperature to control how greedy we want to learn from the embedding-based prediction. We then train BERT on the pseudo training sentences by minimizing the cross entropy H between the sharpened embedding-based prediction $\mathbf{l}$ and the actual output prediction $\mathbf{y}$ of BERT, namely

$$H(x) = \sum_x \sum_i l_{xi} \log \frac{l_{xi}}{y_{xi}}. \tag{5}$$

Note that it is possible to further improve the performance by substituting the language model with powerful counterparts (e.g., RoBERTa-large) or incorporating joint aspect-sentiment analysis [36]. We are also aware that more advanced methods have been proposed for aspect and sentiment extraction [21, 23, 31]. In this work, we only employ a basic setting because the design of classifier is beyond the major focus of our work.

4.2.3 Selecting Aspect-specific Phrases via Robust Model Ensemble. The goal of this step is to strengthen our previous sentence-level model from two angles: (1) Since the phrase extraction step will inevitably introduce non-aspect 'background" phrases, the classifier is expected to cross out them before clustering. However, in the previous step, the model is only trained on aspect-indicative sentences, thus may not perform as well in excluding background phrases. (2) As explained above, the previous model only benefits from clean and high quality sentence-level signals distilled from the embedding space. In contrast, the remaining corpus remain unexplored, which may still contain rich semantic knowledge not captured by the embedding space. To solve these problems, we propose to further enhance the language model by ensembling it with the embedding space for phrase-level aspect classification. Different with previous step, we additionally require the model to identify background phrases by training BERT to output a uniform aspect distribution on them. We also adopt a self-training strategy for the classifier to learn from the entire corpus by sharpening its phrase prediction. Specifically, to enhance the robustness of self-training, we select high-quality training phrases by exploiting the wisdom from both models (i.e., the opinion-oriented embedding and fine-tuned BERT). We take phrases with high output probability from both models as pseudo training data, and generate pseudo labels l'_s for phrase s as:

$$l'_{si} = \begin{cases} \frac{\exp(\alpha y_{si})}{\sum_{i \in |A|} \exp(\alpha y_{si})} & y_{si} \geq \theta_1, \overline{\mathbf{w}}_s^T \mathbf{a}_i \geq \theta_2 \\ \frac{1}{|A|} & \forall i \in |A|, y_{si} < \theta_1, \overline{\mathbf{w}}_s^T \mathbf{a}_i < \theta_2 \end{cases}, \tag{6}$$

where $\overline{\mathbf{w}}_s$ is the averaged embeddings of words in the phrase s, y_{si} is the predicted probability from BERT, and $\overline{\mathbf{w}}_s^T \mathbf{a}_i$ is the directional similarity between phrase s and the i^{th} aspect in the embedding space. θ_1 and θ_2 are two probability thresholds. To explain, if both models output high probability for the same aspect, we count it as positive training samples. Likewise, negative samples are added when both models output low probability for any aspect. Through robust self-training on newly-encountered phrases, the model learns knowledge from the entire unlabeled corpus, which complements with the clean sentence-level signals it received in the previous step. During inference, we use the robustly fine-tuned BERT as the final classifier. The same threshold θ_2 is used to cross out background phrases.

4.3 Fine-Grained Opinion Mining

Given predictions from the previous stage, we can already organize opinions according to their aspects and sentiments. However, phrases located in the same aspect and sentiment may still cover diverse and heterogeneous opinions, varied by their subjects and targeted characteristics. Thus, we propose to mine fine-grained opinions by automatically forming clusters under each aspect and sentiment. To guarantee that phrases belonging to the same cluster convey consistent and coherent meanings, we require them to locate near each other in the semantic space. We achieve this by clustering in the fine-tuned BERT embedding space, as it explicitly encodes discriminative semantics after training and fine-tuning in the previous two stages.

Specifically, we adopt the bottom-up hierarchical agglomerative clustering [7], which treats each phrase as a singleton cluster at the outset, and then successively merges pairs of phrases until the euclidean distance between phrases in the same cluster exceeds a pre-defined threshold T_c. The final output of *FineSum* is the fine-grained clusters under each aspect and sentiment. The granularity of clusters can also be flexibly adjusted via the threshold T_c without repetitively running the phrase extraction and classification step. In real application, the best value of T_c should be adjusted to suit the desired granularity, which is beyond the scope of this paper.

5 EXPERIMENTS

As no off-the-shelf evaluation framework exists, we evaluate model performance on two major tasks in our framework: (i) opinion phrase classification and (ii) fine-grained opinion clustering. We create extra human annotations for qualitative analysis only.

5.1 Datasets and Experimental Settings

5.1.1 Datasets. We experiment on restaurant and laptop reviews. Details of dataset statistics can be found in Table 1. We additionally provide four seed keywords for each aspect and sentiment as weak supervision. We show an example in Table 2 Keywords are selected by following previous work [14]. However, we display the robustness of our methods on seed keywords in Sec. 5.4.

Restaurant: We collect reviews from the Yelp Dataset Challenge[4] as our in-domain training and summarization corpus. We gather reviews from 42 businesses, where each business has at least 100 reviews and includes the keyword "restaurant" in its meta business type list. The average number of reviews for each restaurant is 599. For evaluation, we use an external benchmark dataset in the restaurant domain of SemEval-2016 [32], which provides sentence-level aspect category and sentiment polarity of each review. Following Huang et al. [14], we remove sentences with multiple labels or with a neutral sentiment polarity to simplify the problem.

[4]https://www.kaggle.com/yelp-dataset/yelp-dataset

	# Training Sentences	# Training Phrases	# Test Reviews
Restaurant	10,000	297,210	643
	Aspects		**Sentiments**
	location, drinks, food, ambience, service		good, bad
	# Training Sentences	**# Training Phrases**	**# Test Reviews**
Laptop	16,000	83,540	307
	Aspects		**Sentiments**
	support, os, display, battery, company, mouse, software, keyboard		good, bad

Table 1: Dataset Statistics.

Restaurant Aspect	Location	street block river avenue
	Drinks	beverage wines cocktail sake
	Food	spicy sushi pizza taste
	Ambience	atmosphere room seating environment
	Service	tips manager waitress servers

Table 2: Example keywords of aspects on Restaurant.

Model	Restaurant-phrase			
	Acc.	Pre.	Rec.	macro-F1
CosSim	55.97±2.29	56.22±1.64	53.64±2.22	53.84±1.67
W2VLDA*	63.34±5.39	61.64±4.80	61.37±5.38	60.93±5.47
BERT	72.39±4.00	63.54±3.86	74.13±3.77	67.55±4.29
JASen*	82.20±3.46	85.32±3.13	79.70±2.34	79.64±3.08
w/o BERT	63.80±2.76	62.76±3.29	60.19±3.34	58.21±2.30
w/o joint	76.82±3.79	66.41±3.75	78.98±3.92	70.43±3.86
FineSum	**88.60±3.58**	**85.44±3.64**	**86.98±3.40**	**84.57±4.23**

Table 3: Quantitative evaluation of restaurant aspect identification on phrase level task. * denotes that the model learns aspect and sentiment jointly.

Model	Restaurant				Laptop			
	aspect		sentiment		aspect		sentiment	
	Acc.	macro-F1	Acc.	macro-F1	Acc.	macro-F1	Acc.	macro-F1
CosSim	64.20	49.85	70.19	63.87	55.39	54.33	70.03	70.84
W2VLDA*	71.05	54.30	77.36	70.30	66.21	65.02	71.69	71.66
BERT	75.98	60.83	79.50	77.85	69.12	67.38	71.24	71.17
JASen*	84.99	**74.85**	83.64	81.74	73.02	72.19	76.47	76.53
w/o BERT	81.18	64.52	67.19	51.96	67.34	67.43	68.12	68.33
w/o joint	84.91	68.89	84.76	84.10	71.01	70.42	78.83	78.82
FineSum	**87.67**	69.90	**86.16**	**84.86**	**76.98**	**75.93**	**79.15**	**79.15**

Table 4: Quantitative evaluation of aspect identification and sentiment polarity on sentence level tasks.

Laptop: We collect reviews of 50 laptop products from the Laptop domain of Amazon Review Dataset [28]. Each product has at least 100 reviews. We also use the laptop domain of SemEval-2016 for evaluation and remove sentences with multiple labels .

5.1.2 Experimental Settings. For text prepossessing, we use NLTK tokenizers and the Stanford CoreNLP [22] parser[5,6]. We set the probability thresholds $\theta_1 = 0.35$, $\theta_2 = 0.30$ because it shows the best performances on our self-annotated validation set. We set the context window length $h = 5$, embedding dimension = 100 by following the settings in the hierarchical topic mining paper [27]. We adopt the default learning rate = $1e-5$ in the Hugging Face transformer package[7], and set batch size = 64 according to the maximum capacity of our computing facilities. When selecting the top-K sentences as training data, we initially set $K = 2000$ and observed promising results, so we did not try a larger value of K. We train and fine-tune BERT with AdamW optimizer for only one epoch to avoid overfitting on noisy pseudo labels. We conducted all model training using a single NVIDIA GTX 1080 Ti.

[5]https://stanfordnlp.github.io/CoreNLP/parse.html
[6]https://stanfordnlp.github.io/CoreNLP/depparse.html
[7]https://huggingface.co/transformers/

5.2 Opinion Classification

5.2.1 Evaluation Details. We evaluate the classifier on sentence-level aspect identification and sentiment polarization tasks using the benchmark test set. However, a strong sentence-level classifier may not perform as well on phrase-level task, so we additionally evaluate the model on phrase-level aspect classification task. Due to the shortage of phrase-level annotation, we manually collect data by randomly sampling 500 extracted phrases from the Restaurant domain.

We compare our approach with a series of weakly-supervised baselines and two variants of our own approach.

- **CosSim**: a Word2Vec embedding based model, which classifies according to the cosine similarity between the averaged word embedding and topic vectors. Topic vectors are calculated as the average of seed keywords.
- **W2VLDA** [12]: an aspect-based sentiment analysis model, which leverages aspect and sentiment keywords as seeds to perform joint topic modeling.
- **BERT** [5]: a language model fine-tuning model. We incorporate sentences containing seed keywords as pseudo training samples to fine-tune BERT-base.
- **JASen** [14]: a state-of-the-art aspect based sentiment analysis model. It first learns aspect-sentiment joint word embedding, then generalizes word knowledge to neural models through weakly-supervised training and iterative self-training.
- **w/o BERT**: A context-free ablation of our model. It uses the opinion-oriented spherical embedding as the only classifier.
- **w/o joint**: A BERT-based ablation without robust ensemble (but fine-tuned on sentence-level classification).

The standard Accuracy, Precision, Recall and macro-F1 are used as evaluation metrics. We run experiments for 5 times and report average performances. For each method, we set a lowest threshold for their output classification probability. We classify phrases with a probability larger than the threshold into the corresponding aspect, otherwise the "none" aspect. The threshold is selected when the model performs best on the validation set. Phrase- and sentence-level evaluation results are shown in Tables 3 and 4, respectively. We have the following observations:

Phrase Classification. (1) Comparing the two tables, we observe that most baselines perform worse on phrases than on sentences due to their different class distribution. Take aspect classification as an example, customized sentence-level benchmark contains only aspect-specific samples, while our self-annotated real-world phrases includes a large portion of background samples. The included noise poses challenges to methods developed under a noise-free assumption. This finding highlights the necessity of noise reduction, which corresponds to the robust ensemble step in FineSum. (2) Despite the presence of noisy phrases, *FineSum* still outperforms other baselines by a large margin. Benefiting from both word-level and contextualized knowledge, it achieves better performances than

Spherical Embedding	Vanilla BERT	BERT in FineSum
allergic reactions	severe allergies	severe allergies
severe allergies	broken leg	have intestinal gastro issues
only mild reactions	severe food poisoning	severe food poisoning
no allergic reactions	go through severe migraine	the allergies drive me crazy
their service was a bit lacking	impeccable service	service is nice
the service was a bit too fast	impeccable food quality	attentive and friendly server
the service was a bit slow	service is impressive	they were nice and attentive
that server was a bit opinionated	literally impeccable star service	Mary's service was informative

Table 5: Qualitative evaluation of fine-grained clustering. We compare opinion clusters from different embeddings that show similar semantics. Conflicting and irrelevant phrases are denoted with red and blue respectively.

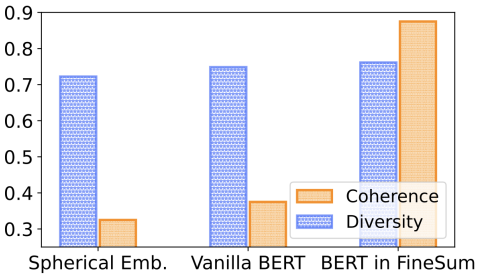


Figure 5: Quantitative evaluation of fine-grained clustering on Restaurant. Higher scores indicates better performances.

models solely based on topic modeling (*W2VLDA*), language model (*BERT*), or word embedding (*CosSim*, *JASen*). (3) *FineSum* also outperforms its two variants. The *w/o BERT* method relies solely on the spherical text embedding, thus it may not utilize rich context information. The *w/o joint* method is fine-tuned only on sentence classification task. Without the phrase-level robust ensemble, its capacity to identify background phrases is limited. As shown in the table, ensemble brings a significant improvement to performance (15.3%↑ Acc. and 20.1%↑ $_{\text{macro}}$-F1), which proves the effectiveness of enhancing confident prediction and excluding noisy one.

Sentence Classification. FineSum also achieves competitive performances on sentence-level classification tasks. One exception is that FineSum shows relatively lower macro-F1 and precision than JASen on restaurant aspect identification. This happens because JASen takes advantage of the mutual information in aspects and sentiments. However, FineSum views them as independent for simplicity. In fact, we can extend FineSum to the joint setting by modifying the word embedding approach, which may lead to even better performance.

5.3 Fine-Grained Opinion Mining

In this section, we empirically evaluate the quality of mined opinion clusters. To illustrate the effectiveness of leveraging fine-tuned BERT embedding for clustering, we compare it with two ablations: (i) Opinion-oriented *Spherical Embedding* introduced in Sec. 4.2 and (ii) Last layer outputs from *Vanilla BERT*, which are not trained on the aspect classification task.

Quantitative Evaluation. We first evaluate the *coherence* and *diversity* of generated clusters quantitatively. *Coherence* measures the semantic consistency of phrases within the same cluster, whereas *diversity* measures how their expressions differ from each other. In principle, a cluster of high quality should be in both high *coherence* and *diversity*, indicating that the model can gather phrases with similar semantic meanings regardless of their expression forms.

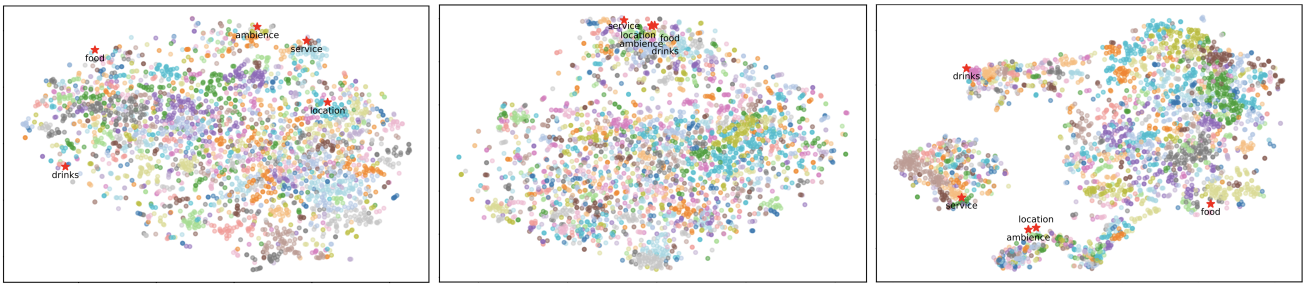
We calculate the two metrics as follows: (i) ***Coherence:*** Given an opinion cluster, we inject an intrusion phrase that is randomly chosen from another cluster. Then we ask annotators to identify the intruded phrase and compute the ratio of correctly identified intrusion instances as the coherence score. Empirically, we observe that phrases within the same cluster usually share common words, making it easy to identify the intruded phrase. Hence, we require the intruded phrase to have at least one overlapped word with other phrases. (ii) ***Diversity:*** The percentage of unique words in each cluster.

Fig. 5 showcases that *BERT in FineSum* significantly outperforms the other two ablations on *coherence*, validating that the fine-tuned BERT embeddings generate semantically coherent clusters. This finding indicates that fine-tuning not only benefits aspect classification, but also leads the model to better distinguish fine-grained sub-aspects within each aspect. Moreover, *BERT in FineSum* achieves slightly higher *diversity* score than the other two methods, indicating that the high coherence score of our approach is not brought by simply gathering phrases with similar words.

Qualitative Evaluation. To intuitively understand the difference between embedding methods, we display semantically similar clusters from them in Table 5. We observe that *Spherical Embedding* relies excessively on overlapped surface words. It tends to gather look-alike phrases even if their meanings are converse. Besides, we also find that *Vanilla BERT*, although less reliant on overlapped words, sometimes suffers from semantic drifts. On the contrary, *BERT in FineSum* forms clusters that are both coherent in meaning and diverse in expression. We further visualize phrase distribution in Fig. 6 to intuitively understand how clusters distribute in the fine-tuned BERT embedding space. Compared with *Spherical Embedding* and *Vanilla BERT*, *BERT in FineSum* displays a clearer and better-separated cluster space. Meanwhile, different aspect names locate far from each other, confirming that the cluster space of *BERT in FineSum* is aspect-distinctive, which fits our goal to learn fine-grained clusters under each aspect.

5.4 Parameter Study

To investigate whether FineSum is sensitive to different choices of seed keywords, we initiate the opinion-oriented embedding with



(a) Spherical Embedding (b) Vanilla BERT (c) Fine-tuned BERT in FineSum

Figure 6: Visualization of opinion clusters on Restaurant. Phrases assigned to the same cluster are denoted with the same color.

Domain: Restaurant	Good	delicious dark ***chocolate*** dipping sauce; it was a good ratio of ***chocolate*** to vanilla; yummy ***chocolate*** brownie; just ok ***chocolate*** berry dessert; amazing ***chocolate*** pudding dessert
		this is a fantastic change while keeping the integrity of so many familiar ***ingredients***; all of the ***ingredients*** went really well together; appreciate how all the ***ingredients*** come together; ***everything*** was well seasoned
Aspect: Food	Bad	left the rest of the ***fish*** untouched; the ***salmon*** was hastily created; the ***fish*** had a weird texture; the ***fish*** seemed a little oily; the ***fish*** was n't that large; i hate raw ***fish***; awful raw ***fish***
		the ***dish*** disappointed my companion; the ***food*** was absolutely underwhelming; they were really disappointed with the ***dish***; my fault for ordering their ***food***, not theirs for making it; I wasn't as impressed with the ***food***

Table 6: A case study of the final output from *FineSum* on Restaurant. We manually *italicize* and bold fine-grained sub-aspects, and highlight sentiment-indicating words with red (good) and blue (bad).

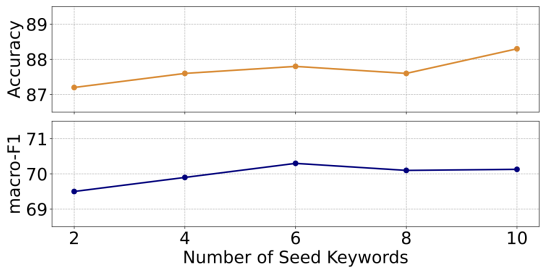


Figure 7: Influence of seed keywords.

different numbers of seeds. Figure 7 shows sentence-level classification results on the Restaurant dataset. As can be observed from the figure, the classification accuracy and macro-F1 remain relatively stable when we alter the number of seeds. This result indicates that the opinion-oriented embedding can learn well-separated semantic space with little human guidance, which validates the robustness of FineSum and opens up possibilities to apply it to diverse domains in the future.

5.5 Case Study

Table 6 shows an example of our system output. We observe that different opinion phrase clusters are well-separated by their aspects and sentiments, which is guaranteed by the opinion phrase classification stage. Probing into each aspect and sentiment, we discover that the model automatically forms clusters which represent concrete sub-aspects or describes particular traits. For example, under the aspect-sentiment pair ⟨food, bad⟩, we find one cluster expressing overall disappointment for their food and two clusters complaining about specific food types. The coherent and meaningful clusters under each aspect validate the effectiveness of clustering with fine-tuned BERT embedding.

6 CONCLUSION

In this paper we propose *FineSum*, a minimally supervised approach for target-oriented, fine-grained opinion summarization. *FineSum* works by first extracting candidate phrases, then classifying them into aspects and sentiments using the opinion-oriented spherical embedding and the weakly-supervised BERT. We further propose to aggregate similar phrases using the fine-tuned BERT embedding to obtain fine-grained opinion clusters. Comprehensive automatic and human evaluation demonstrate that our approach generates high-quality phrase-level summarization.

ACKNOWLEDGEMENTS

Research was supported in part by US DARPA KAIROS Program No. FA8750-19-2-1004 and INCAS Program No. HR001121C0165, National Science Foundation IIS-19-56151, IIS-17-41317, and IIS 17-04532, and the Molecule Maker Lab Institute: An AI Research Institutes program supported by NSF under Award No. 2019897, and the Institute for Geospatial Understanding through an Integrative Discovery Environment (I-GUIDE) by NSF under Award No. 2118329. Any opinions, findings, and conclusions or recommendations expressed herein are those of the authors and do not necessarily represent the views, either expressed or implied, of DARPA or the U.S. Government.

REFERENCES

[1] Reinald Kim Amplayo and Mirella Lapata. 2021. Informative and Controllable Opinion Summarization. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. 2662–2672.

[2] Stefanos Angelidis, Reinald Kim Amplayo, Yoshihiko Suhara, Xiaolan Wang, and Mirella Lapata. 2021. Extractive Opinion Summarization in Quantized Transformer Spaces. *Transactions of the Association for Computational Linguistics* 9 (2021), 277–293.

[3] 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*. 3675–3686.

[4] Eric Chu and Peter Liu. 2019. MeanSum: a neural model for unsupervised multi-document abstractive summarization. In *International Conference on Machine Learning*. PMLR, 1223–1232.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 4171–4186.

[6] Giuseppe Di Fabbrizio, Amanda Stent, and Robert Gaizauskas. 2014. A hybrid approach to multi-document summarization of opinions in reviews. In *Proceedings of the 8th International Natural Language Generation Conference (INLG)*. 54–63.

[7] Werner Dubitzky, Olaf Wolkenhauer, Hiroki Yokota, and Kwang-Hyun Cho. 2013. *Encyclopedia of systems biology*. Springer Publishing Company, Incorporated.

[8] Ahmed El-Kishky, Yanglei Song, Chi Wang, Clare R Voss, and Jiawei Han. 2014. Scalable Topical Phrase Mining from Text Corpora. *Proceedings of the VLDB Endowment* 8, 3 (2014).

[9] Hady Elsahar, Maximin Coavoux, Jos Rozen, and Matthias Gallé. 2021. Self-Supervised and Controlled Multi-Document Opinion Summarization. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. 1646–1662.

[10] Lea Frermann and Alexandre Klementiev. 2019. Inducing Document Structure for Aspect-based Summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 6263–6273. https://doi.org/10.18653/v1/P19-1630

[11] Kavita Ganesan, ChengXiang Zhai, and Jiawei Han. 2010. Opinosis: A Graph Based Approach to Abstractive Summarization of Highly Redundant Opinions. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*.

[12] Aitor García-Pablos, Montse Cuadros, and German Rigau. 2018. W2VLDA: almost unsupervised system for aspect based sentiment analysis. *Expert Systems with Applications* 91 (2018), 127–137.

[13] Xiaotao Gu, Zihan Wang, Zhenyu Bi, Yu Meng, Liyuan Liu, Jiawei Han, and Jingbo Shang. 2021. UCPhrase: Unsupervised Context-aware Quality Phrase Tagging. *arXiv preprint arXiv:2105.14078* (2021).

[14] Jiaxin Huang, Yu Meng, Fang Guo, Heng Ji, and Jiawei Han. 2020. Aspect-Based Sentiment Analysis by Aspect-Sentiment Joint Embedding. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 6989–6999.

[15] Luyang Huang, Lingfei Wu, and Lu Wang. 2020. Knowledge Graph-Augmented Abstractive Summarization with Semantic-Driven Cloze Reward. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 5094–5107.

[16] Jinbae Im, Moonki Kim, Hoyeop Lee, Hyunsouk Cho, and Sehee Chung. 2021. Self-Supervised Multimodal Opinion Summarization. *arXiv preprint arXiv:2105.13135* (2021).

[17] Masaru Isonuma, Junichiro Mori, Danushka Bollegala, and Ichiro Sakata. 2021. Unsupervised Abstractive Opinion Summarization by Generating Sentences with Tree-Structured Topic Guidance. *arXiv preprint arXiv:2106.08007* (2021).

[18] Kundan Krishna and Balaji Vasan Srinivasan. 2018. Generating topic-oriented summaries using neural attention. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. 1697–1705.

[19] Lun-Wei Ku, Yu-Ting Liang, Hsin-Hsi Chen, et al. 2006. Opinion extraction, summarization and tracking in news and blog corpora.. In *AAAI spring symposium: Computational approaches to analyzing weblogs*, Vol. 100107. 1–167.

[20] Sachin Kumar and Yulia Tsvetkov. 2018. Von mises-fisher loss for training sequence to sequence models with continuous outputs. *arXiv preprint arXiv:1812.04616* (2018).

[21] Chengxi Li, Feiyu Gao, Jiajun Bu, Lu Xu, Xiang Chen, Yu Gu, Zirui Shao, Qi Zheng, Ningyu Zhang, Yongpan Wang, et al. 2021. SentiPrompt: Sentiment Knowledge Enhanced Prompt-Tuning for Aspect-Based Sentiment Analysis. *arXiv preprint arXiv:2109.08306* (2021).

[22] Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*. 55–60.

[23] Yue Mao, Yi Shen, Chao Yu, and Longjun Cai. 2021. A joint training dual-mrc framework for aspect based sentiment analysis. *arXiv preprint arXiv:2101.00816* (2021).

[24] Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On Faithfulness and Factuality in Abstractive Summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 1906–1919.

[25] Yu Meng, Jiaxin Huang, Guangyuan Wang, Zihan Wang, Chao Zhang, Yu Zhang, and Jiawei Han. 2020. Discriminative topic mining via category-name guided text embedding. In *Proceedings of The Web Conference 2020*. 2121–2132.

[26] Yu Meng, Jiaxin Huang, Guangyuan Wang, Chao Zhang, Honglei Zhuang, Lance Kaplan, and Jiawei Han. 2019. Spherical text embedding. *Advances in Neural Information Processing Systems* 32 (2019), 8208–8217.

[27] Yu Meng, Yunyi Zhang, Jiaxin Huang, Yu Zhang, Chao Zhang, and Jiawei Han. 2020. Hierarchical topic mining via joint spherical tree and text embedding. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*. 1908–1917.

[28] Jianmo Ni, Jiacheng Li, and Julian McAuley. 2019. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. 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)*. 188–197.

[29] Nadav Oved and Ran Levy. 2021. PASS: Perturb-and-Select Summarizer for Product 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)*. 351–365.

[30] Michael Paul, ChengXiang Zhai, and Roxana Girju. 2010. Summarizing contrastive viewpoints in opinionated text. In *Proceedings of the 2010 conference on empirical methods in natural language processing*. 66–76.

[31] Haiyun Peng, Lu Xu, Lidong Bing, Fei Huang, Wei Lu, and Luo Si. 2020. Knowing what, how and why: A near complete solution for aspect-based sentiment analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 8600–8607.

[32] 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)*. Association for Computational Linguistics, San Diego, California, 19–30. https://doi.org/10.18653/v1/S16-1002

[33] Jingbo Shang, Jialu Liu, Meng Jiang, Xiang Ren, Clare R Voss, and Jiawei Han. 2018. Automated phrase mining from massive text corpora. *IEEE Transactions on Knowledge and Data Engineering* 30, 10 (2018), 1825–1837.

[34] Jiaming Shen, Yunyi Zhang, Heng Ji, and Jiawei Han. 2021. Corpus-based open-domain event type induction. *arXiv preprint arXiv:2109.03322* (2021).

[35] 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*. 5789–5798.

[36] 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*, Vol. 34. 9122–9129.

[37] Zheng Zhao, Shay B Cohen, and Bonnie Webber. 2020. Reducing quantity hallucinations in abstractive summarization. *Findings of the Association for Computational Linguistics: EMNLP 2020* (2020).