A Multi-Granularity Opinion Summarization Method

Anonymous ACL submission

Abstract

Existing opinion mining (OM) is limited to applications on commercial reviews, with *aspect* and *sentiment* of the opinions in a coarse-grained form. In this paper, we further explore the definition of OM by extending the concepts of *aspect* and *sentiment*, and propose an opinion summarization method based on **M**ulti-granularity **C**lustering and **B**ERT (Jacob et al., 2018), i.e., MCB for emergent online discussion record in keeping with the further definition. A supporting Chinese corpus, **ZH45** comprising 45 groups of discussion, and assorted metrics are also proposed. Experiments based on **ZH45** and the metrics demonstrate that MCB produces succinct and insightful opinion summaries.

1 Introduction

Opinion mining (OM), usually interchangeable with the term *sentiment analysis*, is a typical task in the field of natural language processing.

In terms of the opinions obtained, the studies of OM can be divided into two types. The one is aspect-based sentiment analysis (ABSA), which aims at generating opinions in the form of triples like *(aspect, opinion, sentiment polarity)*; the other is opinion summarization (OS) aggregating the opinions in textual form. The advantage of ABSA is the structured output, which is conducive to post-processing, while OS is more informative and readable. OM plays a fantastic role in the big data era. With the mainstream opinions obtained, people can straightforwardly grasp overall cognition, then make decisions about the object of concern without browsing every piece of information.

Despite the considerable advantages, the full development of OM is still far due to two limitations: (1) The premature technical system. In existing researches, aspects of the studied entity are always *predictable* and *concrete* attributes (e.g., the hygiene and service for hotel, the wine, flavor, and price for restaurant, etc.). Similarly, the sentiments or opinions are limited to *simple description*, which can be interpreted as adjectives, or even sentiment polarities. The definition attenuates the difficulty but also the versatility of OM. (2) The supporting evaluation system and corpora have yet to be complemented. The assessment of OS has to mirror that of automatic summarization, adopting metrics such as BLEU (Kishore et al., 2002) and ROUGE (Lin, 2004). The rigorous supervised metrics are not suitable for semi-supervised or unsupervised summarizers, as they tend to underestimate the opinion summaries. Furthermore, the annotation of reference summary is burdensome. Existing quality corpora are all commodity review sets (Chu and Liu, 2019; Bražinskas et al., 2020), preventing OM from penetrating into public opinion monitoring, current affair summary, and big data sentiment analysis.

On a separate note, under the gaze of COVID-19, the face-to-face contact has to be restricted. Consequently, social media like webinars, message boards, micro blogs, etc. have been increasingly spotlighted and used (Jiang et al., 2021), and online discussion record (hereinafter referred to as *discussion*) has been mushrooming all over the social network. The contents of discussion cover quite a board realm, implying substantial exploring value. However, compared to dialogue, news, and commercial comment, discussion possesses excessive volume, relax structure, and miscellaneous expressions, which really impede the progress of its processing.

Given the task framework and the online discussion record above, we propose to further explore the definition of OM by extending the *aspect* and *sentiment* from *concrete*, *knowable* concepts to *abstractive*, *implicit* concepts.

Subsequently, we introduce an OS method based on **M**ulti-granularity **C**lustering and **B**ERT (MCB) for summing up the mainstreams of complicated textual data resembling discussion in an unsupervised, extractive fashion. Through disassembling the text into sentences and terms, MCB bypasses the unstructured problem. At different levels of granularity, MCB employs suitable clustering algorithms in line with the data characteristics and phased needs (e.g., dimensionality or volume reduction, split-flow, and aggregation). In sentence level, we leverage fine-tuned BERT models to inject external knowledge into the framework and advance the exploitation of deep semantics. With the aid of BERT, we add subjectivity analysis and TransfoRank analysis to MCB.

For the sake of assessment, the paper introduces a Chinese corpus, **ZH45** comprising 45 groups of discussion from Zhihu. Zhihu is a large-scale Chinese forum, where objects discussed range from social phenomena to emotional issues. Additionally, a system for evaluation of unsupervised opinion summarizers is proffered. It incorporates the automatic metrics and artificial scoring, evaluating the summarizers from aspects and opinions incrementally. On the basis of our corpus and metrics, experiments including ablation studies and a case study are conducted to verify the practicality and superiority of MCB.

In a nutshell, our contributions include: (1) we comb through the flux of OM, and put up with a deepened conception of the task in the light of emergent data sources, the *discussion*; (2) we propose MCB, an OM method based on multi-granularity clustering and the SOTA language model, BERT for the further task as a baseline, and prove its effectiveness with abundant experiments; (3) we proffer a Chinese corpus consisting of 45 discussions, and an evaluation system for more unsupervised opinion summarizers to refer in the future.

2 Exploration of the concept of OM

2.1 Development of Opinion Mining

The idea of OM was raised in 2003 as "*processing a set of search results for a given item, generating a list of product attributes (quality, features, etc.) and aggregating opinions about each of them (poor, mixed, good)*" (Dave et al., 2003), followed by a technique of classifying the product review sentences according to the sentiment polarity they contain.

From then on, OM was applied to various commercial review sets and short-text social platforms. Statistical analysis on item attributes gained from customers' comments helped the developers to orient themselves in product improvement and information delivery (Claudia and Rachel, 2013; Chen et al., 2014; Sebastiano et al., 2015). As OM evolved, the dimension of emotion had increased. Conrad and Schilder (2007) added subjectivity analysis to polarity analysis when mining opinions from legal blogs. De Choudhury and Counts (2013) used keywords from positive and negative texts as emotion summaries. In the period, some researchers also tried to extract the informative and insightful sentences from the texts to form the opinion summary for certain surveys (Ku et al., 2006; Meng et al., 2012). Ganesan et al. (2010) were the first to state the connection between OM and automatic summarization straightforwardly.

In 2016, SemEval first published the formal definition of ABSA (Pontiki et al., 2016), impelling it to be a relatively complete technical system. The studies mining opinions in tuple form can be categorized under ABSA (Wang et al., 2016, 2017; Tang et al., 2016; Xu et al., 2018; He et al., 2018), whatever algorithms or models were employed. Recently, Xu et al. (2019) post-trained BERT and reached SOTA on multiple ABSA tasks, and Miao et al. (2020) followed the methods of sequence tagging and classification, but significantly reduced the labeled data to achieve SOTA. Researches in OS has also been boosted. For the commercial reviews, researchers have made efforts to create review-like summaries with the most popular opinions extracted (Suhara et al., 2020; Angelidis et al, 2021; Amplayo et al., 2021), or generated by the language models (Kumar et al., 2021). But what for the opinion summary of social media? It is worthwhile devoting more effort to this.

2.2 Further Opinion Mining

This paper aims at bringing the task definition of OM a step further, especially the *source data* and the concepts of the term *aspect* and *sentiment*.

With respect to the data, the crux of discussion processing is explicated as follow:

Multi opinions towards single object. Usually, the thrust of a discussion is not unique. It is inappropriate to preserve the salient contents as in

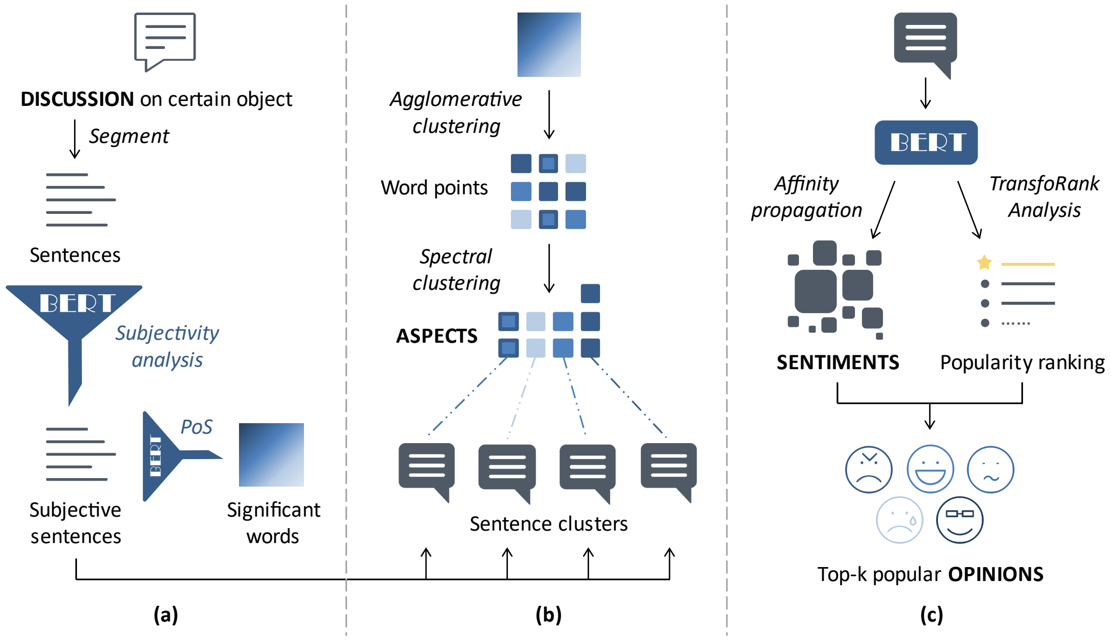

Figure 1: An overview of the OS method based on Multi-grained Clustering and BERT (MCB) in the paper. (a) In preprocessing stage, the fine-tuned BERT models function as filters. (b) In aspect extracting stage, we cluster twice to obtain the extended aspects. (c) In sentiment generalizing stage, affinity propagation and TransfoRank analysis are conducted simultaneously.

other summarization tasks, for some second mainstreams would be omitted.

Excessive volume. Faced with the extensive text, the methods have to cut it down to a manageable size, or be invulnerable to data overload problem.

Relax structure. An individual essay, paragraph or even sentence can become a comment in a discussion, making it difficult to start from the natural comment level as usual.

Miscellaneous expressions. Myriad ironies and degressions are blended in a discussion, hoodwinking the summarizers from figuring out true mainstreams. In addition, similar semantics may hide behind distinct expressions, which leads to the sparse problem.

Existing OM methods may well have difficulties coping with complicated texts, as they extract *aspect* and *sentiment* in the form of *terms*. A single term can hardly represent an aspect of a whole complex, or fully convey an attitude towards something.

The further OM should be able to break the limitations. Inspired by "*opinion mining analyzes people's opinions, appraisals, attitudes, and emotions toward entities, individuals, issues, events, topics, and their attributes*" (Liu, 2011), we suggest selecting such *individuals*, *issues*, *events*, and *topics* as research target, and generalizing *macroscopic angles* instead of tangible attributes as the *aspect*. For the same reason, sentiment is extended to cognitive concept, embracing *insights*, *inferences*, *appraisals*, *attitudes*, and *emotions*. The representation of *extended aspect* or *sentiment* can be defined as a cluster of semantically coherent terms (Zhang et al., 2018), or representative vectors in the semantic space. Subsequently, the representations are collected to form tuples, or help work out summaries.

3 Methodology

3.1 Preprocessing

At the beginning of our method, we trim the relax structure by splitting the given discussion into sentences. Through limiting the length of each sentence and selecting the separators, we try to ensure that every sentence contains one opinion at most, which is different from a linguistic sentence.

Then the subjectivity analysis is implemented by a BERT model. The model is simply fine-tuned on 7500 sentences manually labeled with subjectivity (e. g. 1 for subjective, and 0 for objective). Like general classification tasks, with the hidden state h_0 of the [CLS] token, the subjectivity is calculated as:

$$y^i = softmax(W^i h_0 + b^i) \tag{1}$$

Apart from alleviating the burden brought by the cumbersome texts, we also insist that sentences with stronger subjectivity are qualified for candidate opinions. The subjective sentences are collected into a set denoted as $S = \{s_1, s_2, \dots, s_m\}$.

Another pre-trained ERNIE model (Xiao et al., 2021) is employed for tokenization and part-of-speech (PoS) tagging. Here we mimic the practice of topic modeling, reserving the tokens with specific part-of-speech tags to eschew non-sense words. A stopword list is also used in the process. Consequently, we get the significant word set $W = \{w_1, w_2, \dots, w_n\}$.

3.2 Aspect extracting

The aspect extracting stage identifies several subsets of W to constitute the aforesaid extended aspects of the complex discussed.

After observing the data, we come up with three conditions where words are likely to be relevant.

- Synonyms and antonyms. That means the words are always interchangeable in the text. From another perspective, they would be especially close in semantic space.
- Logical dependencies. For example, *nurse* and *surgery*, *politics* and *economy*, *bloom* and *fruit*, etc. The various dependencies bring the words together, reflecting global statistical characteristics.
- Relevance under particular circumstance. It means that the words are irrelevant usually but relevant within a certain discussion, which corresponds with local statistical characteristics in task-related corpus.

Firstly, we decide to adopt the Word2Vec representations (Mikolov et al., 2013) and the agglomerative clustering algorithm to aggregate the synonyms and the antonyms into *word-points*. The plain algorithm can conveniently control the extent of aggregation by set a similarity threshold. In this paper, we set the threshold to 0.6 empirically and get satisfactory word-point set $P = \{p_1, p_2, \dots, p_o\}$. On the other hand, the short texts are always sparse, noisy and ambiguous (Shi et al, 2018), and the operation mitigates the problems by reducing the dimensions.

Secondly, to give consideration to both global and local relevancy, we compute the frequency $freq_i$ of each word-point in S:

$$S_{p_i} = \{s | w \in p_i, s\ contains\ w, s \in S\} \quad (2)$$

$$freq_i = |S_{p_i}| \quad (3)$$

and the same for co-occurrence frequency $Co_{i,j}$ of every two word-points. Co-occurrence is accessible yet plausible in smaller corpus, especially when the discussion is not large enough to fine-tune the word vectors.

The word-points with low frequencies are removed through a threshold related to the scale of the discussion. Thus, the similarity matrix of remaining word-points can be calculated as:

$$Sim_{i,j} = \frac{pv_i \cdot pv_j \cdot \ln(Co_{i,j} + e)}{\|pv_i\| \|pv_j\|} \quad (4)$$

where pv_i stands for the representation of word-point p_i obtained by averaging the vectors of the words in the word-point. The similarity matrix is the input for spectral clustering (Ng et al., 2002). The graph-based algorithm tallies with the organizational form of the word-points in the discussion. We set the number of clusters between 3 and 6, and the best number is assigned according to the silhouette coefficient. The output clusters of word-points $A = \{a_1, a_2, \dots, a_p\}$ are candidates for the extended aspects, among which we will get some clusters of non-sense words. To weed them out thoroughly, we identify these loose clusters by comparing the intra-cluster and inter-cluster co-occurrence:

$$compact_i = \sum_{p_j \in a_i, p_k \in a_i, j \neq k} Co_{j,k} \quad (5)$$

$$CR_i = \frac{\sum_{n=1}^{p} compact_n - compact_i}{(p-1) \cdot compact_i} \quad (6)$$

where $compact_i$ is the sum of $Co_{j,k}$ of every two word-points in the candidate aspect a_i, and CR_i means the compact degree of a_i. In our method, the loosest cluster will be abandoned if its compact degree is greater than a certain value, and there are enough clusters ($p > 4$ in our study). Then we get the final aspects A^*.

Thirdly, the sentences in S are categorized into the aspects above. The words involved in A^* can vote for the sentences they belong to as:

$$Count_w^s = \begin{cases} 1, & s\ contains\ w \\ 0, & otherwise \end{cases} \quad (7)$$

$$Vote_s^a = \sum_{p_i \in a} \sum_{w \in p_i} Count_w^s freq_i \quad (8)$$

In which $Vote_s^a$ is number of votes for sentence s going to aspect a. In this way, the word-point frequency act as the voting weight. After voting, most sentences are grouped under one or more aspects, while seldom that contain no voter words will be left out.

3.3 Sentiment generalizing

The sentiment generalizing stage is intended to further aggregate the subjective sentences in each group in terms of the emotions they expressed. Given a group under an aspect $S_{a_i} = \{s_1^i, s_2^i, \dots, s_q^i\}$, we first get its embeddings $SV_i = \{sv_1^i, sv_2^i, \dots, sv_q^i\}$. The encoding model is a Chinese BERT model pre-trained with whole word masking (Cui et al., 2021) and fine-tuned on a Chinese natural language inference (NLI) dataset. Although the addition of the model enhances the understanding of deep semantics behind the multiform expressions, note that our method is not tied to any BERT model, including the aforementioned ones for subjectivity analysis and PoS tagging.

In the following, SV_i is fed into two algorithms in parallel. Assuming that sentences under the same topic vary in attitude and emotion most, affinity propagation (AP; Frey and Dueck, 2007) is adopted to generalize the *extended sentiments*. AP excels at clustering multi-class, high-dimensional data, but it has higher complexity than other algorithms. In view of the above, we apply AP to this latter step for the sentiment clusters like $S_{a_i,e_j} = \{s_1^{i,j}, s_2^{i,j}, \dots, s_r^{i,j}\}$ (e for *emotion*). Simultaneously, we put forward TransfoRank by replacing the original similarity function in TextRank (Mihalcea and Tarau, 2004) with a cosine similarity matrix of SV_i, to work out the popularity ranking of S_{a_i}.

We mimic the skill factor in multifactorial evolutionary algorithm (MFEA; Gupta et al., 2015) to design the popularity factor of the sentiment clusters:

$$Pop_{i,j} = \min_{s \in S_{a_i,e_j}} TransfoRank(s) \tag{9}$$

where $TransfoRank(s)$ is the ranking of the sentence by TransfoRank. A smaller popularity factor signifies higher popularity. The central sentences of the K most popular sentiment clusters are extracted for the mainstreams of the aspect. Ultimately, mainstreams coming from all aspects compose the opinion summary.

4 Evaluation of furthered OS

4.1 The ZH45 Corpus

In order to add fuel to the research of further OM, we introduce ZH45, a medium-scale Chinese OM corpus. The ZH45 is constructed on the well-known Q&A community on the Chinese Internet, Zhihu. In the community, the users can pose questions, and discuss other users' questions in turn. Zhihu has a column named "How do you view / evaluate X", where X symbolizes the object discussed, covering social phenomena, news, particular communities, interpersonal problems, and celebrities, etc. The discussion taking place in the column meets the definition of discussion in the paper. We selected and crawled 45 of them. After filtering out the non-text comments, 165K comments were collected in total. The number of comments under each question ranges from hundreds to more than 10K, and the comments vary in length from a few characters to thousands of characters. The corpus contains no reference summaries, which is helpful for unsupervised methods. Actually, the crowdsourcing is unfeasible, as it is unrealistic for people to digest then summary discussions with thousands of comments. The high-quality OS corpus SPACE (Angelidis et al., 2021), where every human-annotated summary is based on 100 reviews, has reached the largest crowdsourcing in the field.

4.2 Metrics

Referring to the work of Angelidis et al. (2021), we evaluate the method from two angles: aspect and opinion summary.

In this paper, the change of the concept of aspect is noteworthy. We believe that there are different appropriate observation angles for different objects, resulting in different aspects. Therefore, we avoid commenting on the right and wrong of the aspects, but focus on their role in the diversion of the subjective sentences.

Aspect Variance is for watching the uniformity of the size of the S_{a_i}. The variance of the sizes of them is taken for the metrics. Considering that between the opinions exist the popularity gaps, a moderate Aspect Variance is acceptable.

Aspect attraction aims at measuring the capacity of the aspects to gather the sentences in a given discussion. We divide the number of subjective sentences collected by A^* by the number of word-points (or words in the case of other methods without word-points) in A for the metric.

Having difficulties in gaining ground-truth annotation in OS research, we are confronted with the challenge of evaluating the existing results without any reference summaries. Human evaluation is indispensable in the situation.

Summary Richness aims at quantizing the information retrieved by the opinion summary. It is computed by preprocessing the summary to get W, and divide the size of W by the number of opinions in the summary.

Summary Coverage examines the ability of the mainstreams in the summary to represent other comments within the discussion. For the metric, we recruited 6 annotators, including undergraduates, graduate students and white-collar workers to carry out a human study. 150 sentences sampled from S arbitrarily were offered with the opinion summary of the discussion, and the annotators had to decide if the sentences were represented by the summary respectively. Each sentence was annotated for 3 times, and there is an average cover rate. We divide the rate also by the number of opinions in the summary, to eliminate the influence of the summary size.

5 Experiments

5.1 Experimental Setup

Dataset. As far as we know, ZH45 is the first corpus to serve further OM, and our experiments take it as the testbed.

Implementation. All of the discussions in ZH45 were involved in evaluation. In the preprocessing stage, the model for subjectivity analysis was `BERT-base-Chinese` (Jacob et al., 2018). As for fine-tuning, we collected 7500 sentences randomly from another 15 posts in the "How do you view / evaluate X" column, and the sentences were annotated by 3 annotators independently as subjective or not. Tokenization and PoS tagging were implemented by a pre-trained multi-task ERNIE model (Xiao et al., 2021) offer in `HanLP` project (He and Choi, 2021). In the sentiment generalizing stage, we borrowed `RoBERTa-wwm-ext` (Cui et al., 2021), and fine-tuned it in the framework of `Siamese-BERT` (Reimers and Gurevych, 2019) using multiple negatives ranking loss. The NLI dataset for fine-tuning was a combination of OCNLI (Hu et al., 2020) and a `Chinese NLI corpus` built on SNLI (Bowman et al., 2015) and MultiNLI (Williams et al., 2018). During training, we used the Adam optimizer, with initial learning rate of 3×10^{-5}. We warmed up the model for the first 10% steps, and ran 5 training epochs in total. The preference in AP clustering was simply set to -1. We let K = 5 while choosing the most popular sentiment clusters. More implementation details agree with the method explained in Section 3.

Aspect		
Methods	Variance	Attraction
MCB	**92270**	**2.9754**
LDA	608132	0.7096
LSI	645968	0.6663
Summary		
Methods	Richness	Coverage
MCB	**5.7339**	0.0270
BERT-Spec	5.1497	0.0214

Table 1: MCB compared with our baselines.

5.2 Results

As for aspect evaluation, noticing the aspect is a collection of words, which is consistent with the concept of the topic, we select LDA (Blei et al., 2003) and LSI (Deerwester et al., 1990) as the baseline models. The upper part of Table 1 shows the Variance and Attraction scores of our method (MCB), LDA, and LSI, which is an encouraging result. Aspects from MCB evidently outperform that from the general topic models in helping produce the opinion summary. From the outputs, LDA and LSI tend to generate highly overlapped word sets, and the sentences are likely to amass around the one with the highest average weight. We conjecture that it is because the discussion has already targeted at a certain object, and the aspects may function as sub-topics under the overall topic, while the typical topic model may concentrate on the latter. Apparently, our method has the ability to discover and distinguish more fine-grained relations.

MCB is the first to break away the commercial reviews and endeavor to solve the further OS task. Hence there is no available baseline. Instead, we took the framework of Jiang et al. (2021) for comparison. For implementations, we encoded the sentences in S by the fine-tuned `RoBERTa-wwm-ext` in Section 4.1, and applied spectral clustering on the embeddings directly. In Euclidean space, the K-nearest neighbors of each cluster center comprised the final summary. The performance of

Aspect		
Methods	Variance	Attraction
MCB	92270	**2.9754**
MCB-Nopnt	129387	2.9384
MCB-Noco	**86072**	2.7817
MCB-FA	1010500	2.7803
Summary		
Methods	Richness	Coverage
MCB	5.7339	0.0270
MCB-Nopnt	5.3315	0.0204
MCB-Noco	5.0480	0.0192
MCB-Noemb	**9.1591**	0.0149
MCB-Spec	5.3335	0.0137

Table 2: Ablation experiment results.

MCB and the baseline (BERT-Spec) is listed in the Table 1 (lower). It can be seen that MCB gains the upper hand with more informative and representative opinion summaries.

5.3 Ablation Study

Since few baseline models can be found, we conducted ablation experiments to confirm the effectiveness of MCB as an extended OS method. The variants considered are as follow:

- MCB-Nopnt: The procedure of aggregating the words into word-points is removed.
- MCB-Noco: The co-occurrence between the word-points is dismissed in the aspect extracting stage.
- MCB-FA: While clustering the word-points, spectral clustering is replaced by agglomerative clustering.
- MCB-Noemb: No BERT models are involved in the sentiment generalizing stage. The top-K popular sentences are worked out by TextRank.
- MCB-Spec: The AP clustering is replaced by spectral clustering in the sentiment generalizing stage.

The MCB-Nopnt, MCB-Noco, and MCB-FA are variants making changes in aspect extracting stage. They are compared with the full method in the aspect evaluation in Table 2 (upper). Among the methods, MCB has a moderate Variance and the highest Attraction. The result of MCB-FA is particularly irrational, and we expelled it from the summary evaluation. In the lower part of Table 2 are some interesting findings. MCB-Noemb wins the highest Richness, followed by our methods. However, MCB maintains the best Coverage far beyond the MCB-Noemb. The reason lies in the summaries themselves: TextRank, the word-frequency-based algorithm, is prone to be misled by sentences that are tediously long and not suitable for mainstreams.

From the above results, it can be concluded that the data dimension reduction, exploitation of statistical characteristics in the task-related data, appropriate clustering algorithms, and BERT models for grasping the deep semantics are all imperative for our method.

MCB
earn money, money, choice, regard… as…, cost, tool {捞钱, 赚钱, 挣钱}, {钱财, 金钱}, {选择}, {当成, 当做, 比作}, {付出, 代价}, {工具}
kid, children, sad, fake smile, smile, happiness {孩子, 孩子们, 小朋友}, {难过, 难受, 心疼}, {假笑, 笑容, 微笑}, {开心, 幸福, 快乐}
joy, society, participate, benefit, pattern, work, develop, capital, popularity {娱乐}, {社会}, {参与, 参加}, {利益}, {方式}, {工作}, {发展}, {资本}, {知名度, 名气}
like, hope, feel, pity, irony {喜爱, 喜欢}, {希望}, {感觉到, 感到, 感觉}, {同情, 怜悯}, {嘲讽, 讥讽, 蔑视}
MCB-Noco
tool, joy, participate, develop, capital, popularity, represent, consume {工具}, {娱乐}, {参与, 参加}, {发展}, {资本}, {知名度, 名气}, {代表}, {消费}
thing, benefit, pattern, work, viewpoint, meaning, long-term {事情}, {利益}, {方式}, {工作}, {看法, 见解, 想法}, {意义}, {长久, 长期}
money, regard… as…, simple, coerce {钱财, 金钱}, {当成, 当做, 比作}, {单纯, 简单}, {强迫, 逼迫, 强行}
sad, like, pity, irony {难过, 难受, 心疼}, {喜爱, 喜欢}, {同情, 怜悯}, {嘲讽, 讥讽, 蔑视}
earn money, compelling, hope, feel, choose {捞钱, 赚钱, 挣钱}, {可笑, 可悲, 可怜}, {希望}, {感觉到, 感到, 感觉}, {选择}
kid, children, fake smile, smile, happiness, brother, family {孩子, 孩子们, 小朋友}, {假笑, 笑容, 微笑}, {开心, 幸福, 快乐}, {弟弟, 家人, 亲人}

Table 3: Aspects of the discussion on the *American "fake smile boy" Gavin*, generated by MCB and MCB-Noco. For brevity, part of word-points with the highest weights in each aspect are shown, with the English meanings labeled above. In every word-point, we display 3 terms at most.

5.4 Case Study

To provide deeper insight into the advantages of our method, we adopt the discussion on the *American "fake smile boy" Gavin* for a case study.

The aspects generated by MCB and the variant with the lowest Variance, MCB-Noco, are shown in Table 3. MCB produces 4 aspects, and we can summarize them as *behavioral intention*, *Gavin himself*, *social effects*, and *public reactions*. MCB-Noco offers 6 aspects, among which 3 aspects are nearly the same with the *social effects* (the first), *public reactions* (the fourth), and *Gavin himself* (the last). The remaining 3 aspects are ambiguous, though.

In Figure 2, we also present the opinion summaries created by MCB and BERT-Spec. with

MCB	通过娱乐他人来赚取钱财并不是一个长久的技能，就我个人而言，这也并不是一个很光彩或是说让人有成就感的事情，感觉会有人说我站着说话不腰疼。(Earning money by entertaining others isn't a long-term policy, it isn't honorable or rewarding for me, either. I feel like some body would say that I don't know the difficulties.) 反正自己选择，自己负责就是了。(As long as he is responsible for his choice.)
	有些人心疼，是因为经历了太多事，希望小孩子可以永远真的开心。(Some feel distressed because they experienced a lot, and hope that the child can be happy forever.) 但凡事都要有个度的，别过度消费孩子的天真可爱，让他们价值观在成年人的世界中受到不好的影响。(Everything has its limits, don't exhaust kids' innocence and loveliness, damaging their values in the adult world.) 希望他能够明白，不是他不可爱了，而是他的可爱不是因为假笑，是因为他自己。(Hope he can know that it isn't that he's not cute, but that he's cute because of himself rather than the fake smile.) 会有些心疼，但是正常吧，普通人也一样啊，为了某些东西，只能强颜欢笑地去做事，要说心疼，那要被心疼的人可多了去了……(Some heartache, but it's normal. Ordinary people also have to struggle with a forced smile for something, so there are many people worth worrying about.)
	有了知名度，就要看自己怎么发展，怎么选择了。(With popularity, it depends on how he develops and chooses.)
	他并没有做错什么，但我认为不妥的是人们对待这件事的态度：这样获得回报的方式不值得我们去追捧。(He is innocent, but I think that people's attitudes are inappropriate: The way of getting a return is not worth pursuing.) 空有悲天悯人的情怀却没有足够大的力量扭转现实，才是作为人间的无奈和悲凉。(Having compassion without enough strength to reverse the reality, that's people's helplessness and sadness.)
Orig.	但凡事都要有个度的，别过度消费孩子的天真可爱，让他们价值观在成年人的世界中受到不好的影响。(Everything has its limits, don't exhaust kids' innocence and loveliness, damaging their values in the adult world.) 有机会得到人们的喜爱，有机会认识不一样的世界，当然是好事啊，只是孩子还小，难免在这个过程中有疑惑，只要父母加以合适的引导与安排，孩子受些辛苦也是值得的。(Having the chance to get people's love and learn about a different world is great of course. The child is still young, however, and he'll have unavoidable doubts in the process. With parents' appropriate guidance and arrangement, it's wo rthwhile for the kid to work hard.)
	会有些心疼，但是正常吧，普通人也一样啊，为了某些东西，只能强颜欢笑地去做事，要说心疼，那要被心疼的人可多了去了……(Some heartache, but it's normal. Ordinary people also have to struggle with a forced smile for something, so there are many people worth worrying about.) 感觉很心疼。(A lot of heartache.)
	他并没有做错什么，但我认为不妥的是人们对待这件事的态度：这样获得回报的方式不值得我们去追捧。(He is innocent, but I think that people's attitudes are inappropriate: The way of getting a return is not worth pursuing.) 这不禁使人陷入深思，这究竟是金钱的扭曲，还是道德的沦丧。(I can't help deeply thinking about whether it's value distortion or moral decay.) 我觉得还是很有道理的。(In my opinion it really makes sense.) 如果他自己感到开心并且喜欢这样的生活，那我也很开心。(If he feels happy and loves such life, I will be happy, too.)

Figure 2: Summaries created by MCB and BERT-Spec (the baseline) for the discussion. The opinions from the same aspect are placed together in one cell.

the second highest Coverage. The two summaries are roughly the same size. In the former summary, we are gratified to find that all of the mainstreams are distinct and insightful, with little overlap between them. Besides, the 4 groups of opinions do match the 4 aspects in contents. The baseline also outputs an acceptable summary, but some over-general opinions (e. g. "*a lot of heartache*" and "*it makes sense*") are blended in, along with an opinion with unknown reference ("*it*" in "*I can't help deeply thinking about…*"), which discounts the quality. As a result, the summary generated by MCB is more informative.

From the comparisons, we see that MCB produces more clear and cohesive aspects for given objects, and the integration of multi-granularity clustering paves a powerful way for OS.

6 Conclusion and Future Works

To review, this paper studies the concept of opinion mining (OM) systematically. On the basis of the existing framework and current application requirements, we come up with an extended task definition for OM, with *aspect* and *sentiment* from concrete, knowable concepts to abstractive, implicit concepts. Instead of simple terms, representations of the aspect and sentiment are turned into clusters of coherent terms or vectors in the semantic space. Without doubt the further OM is much more flexible, and ready to mine more insightful and multifaceted opinions from vaster realm.

We also proposed MCB, an OS method based on multi-granularity clustering and BERT models to bring out our conception into reality. Moreover, a Chinese corpus and a set of evaluation metrics are served for assessment of more summarizers in future. As the experiments demonstrated, our method is well-designed and effective in handling complicated textual data, producing representative, readable opinion summaries rich in information.

One limitation of the current MCB is that the method makes use of the BERT models in a feature-based fashion, and the inner semantic relevance of the task-related data may not be fully exploited. In the future, we will attempt to incorporate graph neural networks (GNN) into OS methods for modeling the extended aspect and sentiment spontaneously. The assisting corpus and metrics will also be perfected in follow-up studies.

References

Reinald K. Amplayo, Stefanos Angelidis, and Mirella Lapata. 2021. Unsupervised opinion summarization with content planning. arXiv preprint arXiv:2012.07808.

Stefanos Angelidis, Reinald K. Amplayo, Yoshihiko Suhara, Xiaolan Wang, and Mirella Lapata. 2021. Extractive opinion summarization in quantized transformer spaces. *Transactions of the Association for Computational Linguistics*, 9:277-293. https://doi.org/10.1162/tacl_a_00366.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993-1022.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pages 632–642. https://doi.org/10.18653/v1/D15-1075.

Arthur Bražinskas, Mirella Lapata, and Ivan Titov. 2020. Unsupervised opinion summarization as copycat-review generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5151–5169. Association for Computational Linguistics. https://doi.org /10.18653/v1/2020.acl-main.461.

Ning Chen, Jialiu Lin, Steven C. H. Hoi, Xiaokui Xiao, and Boshen Zhang. 2014. AR-Miner: Mining informative reviews for developers from mobile app marketplace. In *Proceedings of the 36th international conference on software engineering*, pages 767-778. https://doi.org/10.1145/2568225.2568263.

Eric Chu and Peter Liu. 2019. MeanSum: A neural model for unsupervised multi-document abstractive summarization. In *Proceedings of the 36th International Conference on Machine Learning*, pages 1223–1232.

Iacob Claudia and Harrison Rachel. 2013. Retrieving and analyzing mobile apps feature requests from online reviews. In *2013 10th working conference on mining software repositories (MSR)*, pages 41-44. https://doi.org/10.1109/MSR.2013.6624001.

Jack G. Conrad and Frank Schilder. 2007. Opinion mining in legal blogs. In *Proceedings of the 11th international conference on Artificial intelligence and law*, pages 231-236. https://doi.org/10.1145/1276318.1276363.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, and Ziqing Yang. 2021. Pre-training with whole word masking for Chinese bert. In IEEE Transactions on Audio, Speech and Language Processing, Pages 657-668. Association for Computational Linguistics. https://doi.org/10.1109/TASLP.2021.3124365.

Kushal Dave, Steve Lawrence, and David M. Pennock. 2003. Mining the peanut gallery: opinion extraction and semantic classification of product reviews. In *Proceedings of the Twelfth International World Wide Web Conference (WWW 2003)*. ACM, pages 519–528. https://doi.org/10.1145/775152.775226.

Munmun De Choudhury and Scott Counts. 2013. Understanding affect in the workplace via social media. In *Proceedings of the 2013 conference on Computer supported cooperative work*, pages 303-316. https://doi.org/10.1145/2441776.2441812.

Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391-407. https://doi.org/10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9.

Brendan J. Frey and Delbert Dueck. 2007. Clustering by passing messages between data points. *science* 315 (5814):972-976. https://doi.org/10.1126/science.1136800.

Kavita Ganesan, Chengxiang Zai, and Jiawei Han. 2010. Opinosis: a graph-based approach to abstractive summarization of highly redundant opinions. https://www.ideals.illinois.edu/handle/2142/16949.

Abhishek Gupta, Yew-Soon Ong, and Liang Feng. 2015. Multifactorial evolution: toward evolutionary multitasking. *IEEE Transactions on Evolutionary Computation 20*, pages 343-357. https://doi.org/10.1109/TEVC.2015.2458037.

Ruidan He, Wee-Sun Lee, Hwee-Tou Ng, and Daniel Dahlmeier. 2018. Exploiting document knowledge for aspect-level sentiment classification. arXiv preprint arXiv:1806.04346. http://dx.doi.org/10.18653/v1/P18-2092.

Han He and Jinho D. Choi. 2021. The stem cell hypothesis: dilemma behind multi-task learning with transformer encoders. arXiv preprint arXiv:2109.06939 (2021). http://dx.doi.org/10.18653/v1/2021.emnlp-main.451.

Hai Hu, Kyle Richardson, Liang Xu, Lu Li, Sandra Kuebler, and Larry Moss. 2020. Ocnli: original chinese natural language inference. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3512–3526. http://dx.doi.org/10.18653/v1/2020.findings-emnlp.314.

Devlin Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805. https://arxiv.org/abs/1810.04805.

Han Jiang, Yubin Wang, Songhao Lv, and Zhihua Wei. 2021. An opinion summarization-evaluation system based on pre-trained models. In *International Joint Conference on Rough Sets*. Springer, Cham, pages 225-230. https://doi.org/10.1007/978-3-030-87334-9_19.

Papineni Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311-318. http://dx.doi.org/10.3115/1073083.1073135.

Lun-Wei Ku, Yu-Ting Liang, and Hsin-Hsi Chen. 2006. Opinion extraction, summarization and tracking in news and blog corpora. In *AAAI spring symposium: Computational approaches to analyzing weblogs (Vol. 100107)*, pages 1-167. https://www.aaai.org/Library/Symposia/Spring/2006/ss06-03-020.php.

Akshi Kumar, Simran Seth, Shivam Gupta, and Shivam Maini. 2021. Sentic computing for aspect-based opinion summarization using multi-head attention with feature pooled pointer generator network. *Cognitive computation (2021)*:1-19. https://doi.org/10.1007/s12559-021-09835-8.

Chin-Yew Lin. 2004. Rouge: a package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74-81. https://aclanthology.org/W04-1013/.

Bing Liu. 2011. Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data. Second Edition. Springer. https://doi.org/10.1007/978-3-642-19460-3.

Xinfan Meng, Furu Wei, Xiaohua Liu, Ming Zhou, Sujian Li, and Houfeng Wang. 2012. Entity-centric topic-oriented opinion summarization in Twitter. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 379-387. https://doi.org/10.1145/2339530.2339592.

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. https://doi.org/10.1145/3366423.3380144.

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. https://aclanthology.org/W04-3252/.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781. https://arxiv.org/abs/1301.3781.

Andrew Y. Ng, Michael I. Jordan, and Yair Weiss. 2002. On spectral clustering: analysis and an algorithm. In *Advances in neural information processing systems*, pages 849-856. https://dl.acm.org/doi/abs/10.5555/2980539.2980649.

Maria Pontiki, Dimitrios Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad Al-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée de Clercq, et al. 2016. SemEval-2016 Task 5: Aspect Based Sentiment Analysis. In *International workshop on semantic evaluation*, pages 19-30. https://dx.doi.org/10.18653/v1/S16-1002.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. arXiv preprint arXiv:1908.10084. http://dx.doi.org/10.18653/v1/D19-1410.

Panichella Sebastiano, Andrea Di Sorbo, Emitza Guzman, Corrado A. Visaggio, Gerardo Canfora, and Harald C. Gall. 2015. How can i improve my app? classifying user reviews for software maintenance and evolution. In *2015 IEEE international conference on software maintenance and evolution (ICSME)*, pages 281-290. https://doi.org/10.1109/ICSM.2015.7332474.

Tian Shi, Kyeongpil Kang, Jaegul Choo, and Chandan K. Reddy. 2018. Short-text topic modeling via non-negative matrix factorization enriched with local word-context correlations. In *Proceedings of the 2018 World Wide Web Conference*, pages 1105-1114. https://doi.org/10.1145/3178876.3186009.

Yoshihiko Suhara, Xiaolan Wang, Stefanos Angelidis, and Wang-Chiew Tan. 2020. Opiniondigest: a simple framework for opinion summarization. arXiv preprint arXiv:2005.01901. http://dx.doi.org/10.18653/v1/2020.acl-main.513.

Duyu Tang, Bing Qin, and Ting Liu. 2016. Aspect level sentiment classification with deep memory network. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 214–224. http://dx.doi.org/10.18653/v1/D16-1021.

Wenya Wang, Sinno Jialin Pan, Daniel Dahlmeier, and Xiaokui Xiao. 2016. Recursive neural conditional random fields for aspect-based sentiment analysis. arXiv preprint arXiv:1603.06679. https://arxiv.org/abs/1603.06679.

Wenya Wang, Sinno Jialin Pan, Daniel Dahlmeier, and Xiaokui Xiao. 2017. Coupled multi-layer attentions for co-extraction of aspect and opinion terms. In *Thirty-First AAAI Conference on Artificial Intelligence*. https://ojs.aaai.org/index.php/AAAI/article/view/10974.

Adina Williams, Nikita Nangia, Samuel R. Bowman. 2018. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. arXiv preprint arXiv:1704.05426. http://dx.doi.org/10.18653/v1/N18-1101.

Dongling Xiao, Yu-Kun Li, Han Zhang, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2021. Ernie-gram: pre-training with explicitly n-gram masked language modeling for natural language understanding. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, pages 1702–1715. https://aclanthology.org/2021.naacl-main.136.

Hu Xu, Bing Liu, Lei Shu, and Philip S. Yu. 2018. Double embeddings and cnn-based sequence labeling for aspect extraction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics. http://dx.doi.org/10.18653/v1/P18-2094.

Hu Xu, Bing Liu, Lei Shu, and Philip S. Yu. 2019. Bert post-training for review reading comprehension and aspect-based sentiment analysis. arXiv preprint arXiv:1904.02232. http://dx.doi.org/10.18653/v1/N19-1242.

Chao Zhang, Fangbo Tao, Xiusi Chen, Jiaming Shen, Meng Jiang, Brian Sadler, Michelle Vanni, and Jiawei Han. 2018. TaxoGen: usupervised topic taxonomy construction by adaptive term embedding and clustering. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2701-2709. https://doi.org/10.1145/3219819.3220064.