AX-MABSA: A Framework for Extremely Weakly Supervised Multi-label Aspect Based Sentiment Analysis

Sabyasachi Kamila[1], **Walid Magdy**[1], **Sourav Dutta**[2] and **MingXue Wang**[2]
[1]School of Informatics, University of Edinburgh
[2]Huawei Research Centre, Dublin, Ireland
{skamila,wmagdy}@inf.ed.ac.uk, {sourav.dutta2,wangmingxue1}@huawei.com

Abstract

Aspect Based Sentiment Analysis is a dominant research area with potential applications in social media analytics, business, finance, and health. Prior works in this area are primarily based on supervised methods, with a few techniques using weak supervision limited to predicting a single aspect category per review sentence. In this paper, we present an extremely weakly supervised multi-label *Aspect Category Sentiment Analysis* framework which does not use any labelled data. We only rely on a single word per class as an initial indicative information. We further propose an automatic word selection technique to choose these seed categories and sentiment words. We explore unsupervised language model post-training to improve the overall performance, and propose a multi-label generator model to generate multiple aspect category-sentiment pairs per review sentence. Experiments conducted on four benchmark datasets showcase our method to outperform other weakly supervised baselines by a significant margin.[1]

1 Introduction

Aspect-based sentiment analysis (ABSA) is a well-known sentiment analysis task which provides more fine-grained information than simple sentiment understanding (Liu, 2012). The main goal of ABSA is to find the aspects and its associated sentiment within a given text. While the works on ABSA have expanded in different directions, it has primarily two sub-tasks, *Aspect Term Sentiment Analysis* (ATSA) and *Aspect Category Sentiment Analysis* (ACSA) (Xue and Li, 2018). ATSA consists of different tasks like aspect term extraction (Li et al., 2018; Luo et al., 2019; Li et al., 2020a; Shi et al., 2021), aspect term sentiment classification (He et al., 2018; Chen and Qian, 2019; Hou et al., 2021), opinion term extraction (Dai and Song, 2019; He et al., 2019; Chen and Qian, 2020b), aspect-oriented opinion term extraction (Fan et al., 2019; Wu et al., 2020a), aspect-opinion pair extraction (Zhao et al., 2020), etc. For example, in the sentence "*The sushi is top-notch, the waiter is attentive, but the atmosphere is dull.*", ATSA would extract the aspect terms '*sushi*', '*waiter*' and '*atmosphere*'; opinion terms '*top-notch*', '*attentive*', and '*dull*'; and their associated sentiments '*positive*', '*positive*' and '*negative*'. The other sub-task ACSA aims to find the higher order aspect categories and its associated sentiment from a given text. In the above example, ACSA would detect the categories as '*food*' (as 'pasta' is a type of 'food'), '*service*' and '*ambience*'; and the associated sentiments as '*positive*', '*positive*' and '*negative*'.

Existing research on ABSA is dominated by supervised methods, where labeled training data is provided (Chen et al., 2017; Xue and Li, 2018; Cai et al., 2021; Liu et al., 2021; Xu et al., 2021; Yan et al., 2021). A few works try to solve the problem in a weakly/semi-supervised manner, where a few labelled samples are provided (Wang et al., 2021a). However, there has been a lack of study on ABSA using *unsupervised methods*, i.e., without using any labelled data. A few works also focused on unsupervised aspect term extraction (Shi et al., 2021). However, such works do not deal with the sentiment associated with the aspects. An existing work on weakly supervised ACSA (Huang et al., 2020) only considered a single aspect category per sentence – thus limiting the task to a larger extent.

Motivated by the above, in this work, we present a methodology for *extremely weakly supervised ACSA* task, where we do not need any labelled training samples. We solve both aspect category detection (ACD) and ACSA tasks (on each review sentence) just by using the surface text of aspect category and sentiment. Given N review sentences, C categories of interest and P polarities of interest, the ACD task generates C clusters, while the

[1]Code, data and model are available at https://github.com/sabyasachi-kamila/AX-MABSA

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 6136 - 6147
December 7-11, 2022

ACSA task generates *(c_i, p_j)* tuples where $c_i \in C$, and $p_j \in P$. As in (Wang et al., 2021b), we adopt the representation learning perspective, wherein representing sentences by class names leads to better clustering. We only use the surface text of the class names and unlabelled sentences to get aspect category and sentiment clusters.

However, in clustering, each review sentence would get only one label, thus limiting the task by a substantial extent. To tackle this, we propose *X-MABSA*, a *multi-label generator model* which makes use of dependency parser (Qi et al., 2020) and a similarity-based attention mechanism to generate multiple categories and associated sentiment polarity labels for each review sentence. In addition, we find that sometimes the representative text of aspect categories (provided as input) is not present (or sparse) in the text corpus. This might lead to skewed representation of the classes in our framework and thus degrade performance. Therefore, we present an automatic surface word selection strategy which would represent the class names better. We combine this with our X-MABSA model and denote it as AX-MABSA.

We also showcase that unsupervised post-training of language model on domain specific data significantly improves the sentence representation and thus achieves better results for ACSA tasks. For this, we post-train BERT language model (Devlin et al., 2019) using domain specific unlabelled data. We perform experiments on four different benchmark aspect-based datasets (Pontiki et al., 2014, 2015, 2016; Cheng et al., 2017), and compare with different supervised and weakly supervised baselines. Our main contributions are as follows:

- an extremely weakly supervised method to solve the ACSA task without relying on any labelled data, and using only the class names as the only provided information;
- an automatic surface word selection strategy for choosing a suitable word corresponding to each aspect and sentiment class;
- use of BERT language model post-training on domain specific unlabelled data for semantic representation of review sentences;
- a multi-label generator model which makes use of a dependency parser and a similarity-based attention mechanism for generating multiple aspect-sentiment labels for each sentence; and
- experimental results comparing our architecture with different existing baselines on four benchmark aspect datasets.

2 Related Work

Aspect Based Sentiment Analysis (ABSA) has gained significant attention for a long time, and research has been done in primarily two directions – Aspect Term Sentiment Analysis (ATSA) and Aspect Category Sentiment Analysis (ACSA).

2.1 Aspect Term Sentiment Analysis

Research on ATSA has been in different sub-categories like,

Aspect Term Extraction In this sub-task, aspect terms associated with a category are extracted from a given text. Prior research on this is based on sequence labelling problem (Ma et al., 2019; Li et al., 2020a). Li and Lam (2017) proposed a neural network-based deep multi-task framework with memory network for extracting aspect terms. Xu et al. (2018) presented a double embedding method which uses CNN (LeCun et al., 1995)-based sequence tagging, while Li et al. (2018) considered summary of opinions expressed in text as well as the history of aspect detection for effective aspect term extraction. Chen and Qian (2020a) proposed a soft prototype-based approach with aspect word correlations to improve quality. A few unsupervised methods have tried to improve performance by using traditional topic modelling-based models. Luo et al. (2019) proposed a neural network based unsupervised model which takes sememes for better lexical semantics. Shi et al. (2021) presented a self-supervised method which works on learning aspect embedding on the word embedding space for aspect extraction.

Aspect-level Sentiment Classification In this sub-task, sentiment labels are assigned to each aspect term. Wang et al. (2016); Liu and Zhang (2017); Ma et al. (2017) proposed an attention-based neural network model for aspect-level sentiment classification (ASC). Tay et al. (2018) modelled relationship between words and aspects using LSTM model (Hochreiter and Schmidhuber, 1997) to improve ASC performance. He et al. (2018) showed that document knowledge transfer improved performance of ASC task. Chen and

Qian (2019) proposed a transfer capsule network for transferring knowledge from document-level sentiment classification, while Hou et al. (2021) adopted a dependency tree-based graph neural network to solve the ASC task.

Aspect-oriented Opinion Extraction This task extracts opinion terms associated with aspect terms. Fan et al. (2019) designed a sequence label model which used LSTM (Hochreiter and Schmidhuber, 1997) for aspect-oriented opinion extraction (AOE). Wu et al. (2020a) proposed a tagging scheme for AOE task which uses CNN (LeCun et al., 1995), LSTM (Hochreiter and Schmidhuber, 1997) and BERT (Devlin et al., 2019) for opinion extraction. Wu et al. (2020b) proposed a transfer learning method for transferring knowledge from sentiment classification task to AOE task.

Recent works on ATSA have introduced more sub-tasks like aspect-opinion pair extraction, aspect-sentiment-opinion triplet extraction, aspect-category-opinion-sentiment quadruple extraction, etc. Yan et al. (2021) proposed a BART (Lewis et al., 2020) -based model to solve all ATSA tasks. Cai et al. (2021) introduced a new task called, aspect-category-opinion-sentiment quadruple extraction, a BERT (Devlin et al., 2019)-based model to deal with implicit aspects and opinion terms. Xu et al. (2021) proposed a new span-level method for the aspect-sentiment-opinion triplet extraction.

2.2 Aspect Category Sentiment Analysis

Aspect Category Sentiment Analysis (ACSA) finds aspect categories and their associated sentiments from a text. Research on this has been conducted on both Aspect Category Detection (ACD) and ACSA tasks. Ma et al. (2018) proposed a word attention-based hierarchical model which takes common-sense knowledge for solving ACSA task. Xue and Li (2018) presented a novel CNN (LeCun et al., 1995)-based model for ACSA task. Liang et al. (2019) proposed an encoding scheme which was aspect-guided and able to perform aspect-reconstruction. Sun et al. (2019) constructed an auxiliary text for aspects and reformed the ACSA as a classification task.

Wang et al. (2020) proposed a novel dependency tree-based model and a relational graph attention network for encoding the sentences. Li et al. (2020b) designed a multi-instance framework for multi-label ACSA task. Cai et al. (2020) reformed the task as sentiment-category with a two-layer hierarchy where the higher layer detected the sentiment while the lower layer detected the aspect category. Liang et al. (2021) presented a semi-supervised framework having a beta distribution-based model. The model finds semantically related words from the context of a target aspect. Liu et al. (2021) solved the ACSA task as a text generative method using BART (Lewis et al., 2020). Zhang et al. (2021) presented aspect sentiment quad prediction task where ACSA was formulated as a paraphrase generation task.

Almost all existing works on ACSA are based on supervised methods. In contrast, this work proposes a method for ACSA which does not require any labelled data and relies only on seed text for aspect class names.

3 Proposed Methodology

Our proposed method, *AX-MABSA*, works on the following components: (a) class name-based clustering, (b) unsupervised language model post-training on domain-specific data for better contextual representation of review sentences, (c) a multi-label generator model to generate multiple categories and associated sentiment labels, and (d) automatic class-representative text selection. The overall framework is depicted in Figure 1.

Problem Formulation: We formulate the extremely weakly supervised ACD and ACSA tasks as: Consider as input a review sentence $x = \{x_1, x_2, x_3, ..., x_n\}$ where x_i is the i^{th} word of the sentence and n is the length of the sentence, along with a list of *C* predefined aspect categories. The output for the ACD task is c categories for a sentence where $c \subset C$. For the ACSA task, the output is a list of tuples (c_j, p_k) where c_j is the j^{th} predicted category and p_k is the k^{th} predicted sentiment polarity corresponding to the category c_j. The sentiment polarity p is from the set $s \subset \{positive, negative\}$.

3.1 ACSA Module

As a primary task, we address the aspect detection based on the seed aspect categories provided as input. We adopt the X-Class model as presented in (Wang et al., 2021b) for solving extremely weakly supervised classification tasks majorly on topic modelling datasets. This module involves four stages: (a) word representations, (b) class representation, (c) class-specific document representation, and (d) document-class alignment.

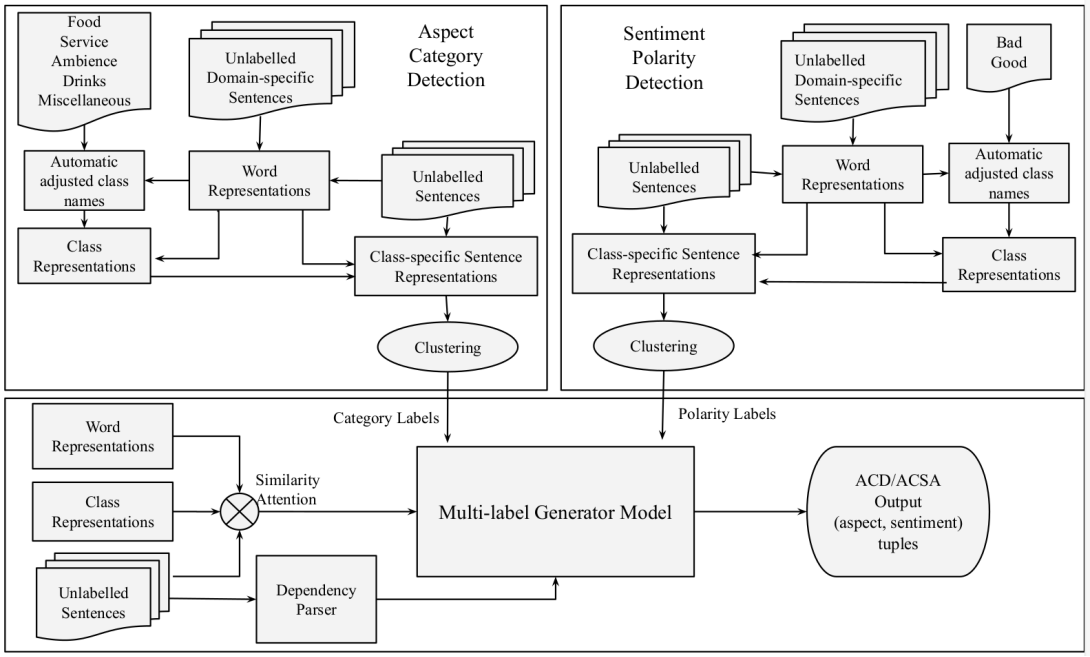

Figure 1: Overview of the proposed AX-MABSA Framework

For word representations, at first, a vocabulary is created from all the input texts. Then, each word's contextual representation is represented using a pre-trained BERT language model (Devlin et al., 2019). The contextual embeddings of each word are averaged to obtain the review sentence encoding, and this representation is denoted as the static word representation s_r.

$$s_r = \frac{\sum_{R_{i,j}=r} z_{i,j}}{\sum_{R_{i,j}=r} 1} \quad (1)$$

Here, $z_{i,j}$ is the contextualized representation and $R_{i,j}$ is the j^{th} word in the review sentence R_i.

As class representation, the representations of the aspect class names are constructed based on the static representations of those words. For example, the category "sports" is represented by the contextual embedding of the word "sports". Then an expansion technique is used to find similar words of each class name words from within the input texts and average those words' contextual representations to obtain the final aspect class embedding.

In class-specific document representation, the representations of the documents or the sentences are guided by the class representations so that the sentences become more aligned to the topics of interest, i.e., the class names. Different attention mechanisms are used over the document representations guided by class representations to get updated document representations. Finally, for document-class alignment, clustering algorithms are used to cluster n-documents to c-clusters (c is the number of classes), wherein the seed class centroids are initialized with the class representations.

Clustering Algorithms: We followed different centroid-based clustering algorithms such as K-Means (Lloyd, 1982), Mini-batch K-Means (Sculley, 2010) and Gaussian Mixture Model (GMM) (Duda et al., 1973); and found that in-general Mini-batch K-Means (mk-means) performs best for the ACD task while GMM performs best for the ACSA task. So, we fix this for our experiments. We used Principal component analysis (Abdi and Williams, 2010) for dimensionality reduction of sentence representation and class representation vectors before clustering. The target dimension is set to 64. We also fixed *random_state* to 42 for centroid initialization. For mk-means, we used batch size 400.

The model requires the surface text of the class names to be present on the dataset for a certain number of times. We feel this is a potential drawback in solving our ACSA task, as some surface text of category names may not be present in the dataset or have a proper meaning representation. For example, the category word "miscellaneous" might have no clear meaning and sometimes might not be present in the dataset. To resolve this issue, we explicitly add the category name to the vocabulary set if it is not found on the dataset. Another drawback of the above approach is that it can only

predict one label per sentence. This is a huge limitation, especially when multiple aspect categories are present in a review sentence. In the following sections, we tackle these issues to propose a robust multi-aspect extraction framework.

3.2 AX-SABSA

We observed that the performance of the implemented ACSA module based on X-Class is poor. One of the reasons is that the words' representations are based on the pre-trained BERT language model (Devlin et al., 2019) which gives more general representations of each word, which works well for topic modelling tasks. However, the aspect terms are more specific to the domains and thus general representation does not provide specific information. Therefore, we suggest that unsupervised post-training of BERT on domain-specific data would lead to better word-representations and thus a better performance.

Unsupervised Post-training of BERT Language Model (UPBERT): We follow a recent model (Gao et al., 2021) which feeds the same input twice, one with dropout masks and the other with different dropout masks, to the encoder. The model optimizes the following objective function:

$$z_i = -log \frac{e^{sim(a_i^{p_i}, a_i^{p_i^{'}})/\gamma}}{\sum_{j=1}^{N} e^{sim(a_i^{p_i}, a_j^{p_j^{'}})/\gamma}} \quad (2)$$

Here, a is hidden state, which is a function of the input sentence and dropout masks p and $p^{'}$.

We feed our collected domain specific unlabeled data of varied sizes to this representation and fine-tune the BERT model. For our experiments, we use *batch size* as 128, *sequence length* as 32, *learning rate* as 3e-5, loss function as *Multiple Negatives Ranking Loss* of the sentence-transformer model (Reimers and Gurevych, 2019). We vary the dataset size starting from 10k samples for training, and get different fine-tuned BERT models corresponding to different data sizes. Finally, we select the fine-tuned model which provides the best performance (using 80k unlabeled training sentences in our case) and apply this UPBERT model for word representation to solve the ACSA task and call our single-label predictor model X-SABSA.

Automatic Class-representative Surface Text Selection Algorithm (ACSSA): Our model suffers when the surface text of class names is not present on the data. Although we add these words to the vocabulary explicitly, their contextual representations become poor. As an immediate solution, we can manually select candidate words corresponding to class names. However, this would be a difficult and tedious job when the number of categories would be high. Also, there can be multiple candidate words for a class name. For example, to represent the category "ambience", one can choose any of the following words: atmosphere, environment, vibes, etc. Similarly, to represent a negative polarity, one can choose any of the following words: bad, problem, pathetic, poor, etc. Depending upon the words we choose, the overall performances vary significantly. So, we propose an algorithm that selects these candidate words automatically given the original class names (see Algorithm 1).

The algorithm ACSSA takes a particular part-of-speech tag (noun for the ACD task, adjective for the ACSA task), dataset, vocabulary and the class names as input and produces a candidate word list as output. Initially, it creates a list *uV* of words from the vocabulary which has a desired part-of-

Algorithm 1 Algorithm for Automatic Class-representative Surface Text Selection

Input: *X* (noun for ACD, adjective for ACSA), dataset *D*, vocabulary *V*, class names *C*
Output: A List *selected* containing Candidate words for each class
Initialize a global array uV[], *T*, targetL[], sourceL[], interL[], goalL[], selected[]
for *w in V* **do**
  $pos \leftarrow getPosTag(w)$
  if *pos == 'X'* **then**
    $uV \leftarrow append(w)$
  end
end
for *i in C* **do**
  $sims \leftarrow findSimilarity(i, uV)$
  $sortedSim \leftarrow argSort(sims)$
  for *j in C* **do**
    if *i!=j* **then**
      $t \leftarrow TopTsimilarwordsfromsortedSim[j]$
      $targetL \leftarrow t$
    end
  end
  $s \leftarrow TopTsimilarwordsfromsortedSim[i]$
  $sourceL \leftarrow s$
  for *w in sourceL* **do**
    if *w not in targetL* **then**
      $interL \leftarrow append(w)$
    end
  end
  $goalL \leftarrow interL$
  $goalL \leftarrow ArgSortOccurance(goalL)$
  $selected \leftarrow firstValue(goaL)$
end

speech tag. It then finds all the similar words for each class name from the list *uV*. We then select *top-T* values from each list. This can be varied depending upon inspection. We fixed it to 10 based on experimental results. Then the similar words for each class are sorted according to the cosine similarity scores. Finally, we sort each list according to their number of occurrences in the dataset. We then select the topmost occurring word from each list as the aspect class representative. This would produce a single candidate word for each class. Thus, the AX-SABSA module uses ACSSA in combination with X-SABSA, to automatically generate better aspect category names.

3.3 AX-MABSA

Since clustering produces only one label for each review sentence, we propose a *Multi-label Generator* model based on dependency parser (Qi et al., 2020) and a similarity-based attention mechanism.

Multi-label Generator Model: This model takes the unlabelled sentences, the sentence representations, the category class representations, and the clustering outputs to generate multiple categories and associated sentiment polarity for each sentence. We illustrate the model using the following example: "The food was good, but it's not worth the wait or the lousy service". The sentence has tags '(food, positive)' and '(service, negative)'.

Parsing the Input Sentences The unlabelled input sentences are parsed by off-the-shelf dependency parser (Qi et al., 2020). The parser outputs a pair of dependencies (word, word[head-1]). The output of the above sentence can be seen in Figure 2. For each word with Noun part-of-speech tag in the sentence, we select those pairs where either the word or word[head-1] is also a Noun. We call these final set of pairs as 'PPairs'. The 'PPairs' for the above sentence are ('food', 'good'), ('wait', 'worth'), and ('service', 'wait'). Observe, in general, the first word in a pair is related to aspects while the second word is associated with sentiment.

Similarity-based Attention We use a similarity-based attention mechanism to assign a desired class label to each of the words in the PPairs. We first obtain the similarity values between the words in the sentence and all the class names using the cosine similarity as: $S_{i,j} = cos(w_i, c_j)$.

Now, we calculate $max_c(S)$ which assigns each word to the highest similar class. For each aspect word in the 'PPairs' if the corresponding $max_c(S)$ is greater than a threshold[2] then we keep those 'PPairs'. We call these filtered pairs as 'FP-Pairs'. Finally, we assign the aspect, sentiment label to each 'FPPairs' based on its corresponding $max_c(S)$ values. If the 'FPPairs' has only one or empty pair, then we consider the clustering outputs as the predicted aspect and sentiment pair.

In the entire setup, we use the UPBERT model mentioned in Section 3.2 for word representation. We refer to the entire model as X-MABSA. When we use the automatic surface word selection algorithm ACSSA in the X-MABSA model, we call that final model AX-MABSA.

Dataset	Rest-14	Rest-15	Rest-16	MAMS
# of Categories	5	5	5	8
# of Sentences	800	582	586	400
Avg # of Aspects/sentence	1.28	1.21	1.18	2.25
Imbalance	5.04	12.10	7.34	9.09

Table 1: Gold Data Statistics. Imbalance value signifies the ratio between the largest and smallest category size.

4 Experimental Setup

We discuss here the datasets we have used, word representations, and different baselines we have selected for our experiments.[3]

4.1 Datasets

We chose the SemEval-2014 restaurant review (Rest-14) (Pontiki et al., 2014), SemEval-2015 restaurant review (Rest-15) (Pontiki et al., 2015), SemEval-2016 restaurant review (Rest-16) and the multi-aspect multi-sentiment (MAMS) (Cheng et al., 2017) datasets for sentence-level aspect category and aspect category sentiment. The Rest-14 data has five categories as food, service, ambience, price, and miscellaneous. Rest-15 and Rest-16 have restaurant, ambience, food, service, and drinks categories. MAMS dataset has food, ambience, price, service, miscellaneous, staff, menu, and place categories. The test data size for all the dataset is reported in Table 1. Imbalance signifies the ratio between the largest class size and smallest class size.

Data for BERT Post-training For BERT post-training, we consider the *Citysearch* data created

[2]We fixed this threshold to 0.45 based on qualitative and quantitative evaluation.

[3]We performed our experiments on Nvidia 1e30 GPUs with CUDA 11. The post-training experiments on an average took 10-15 minutes each, and each model of our proposed method took around 1-2 minutes to run.

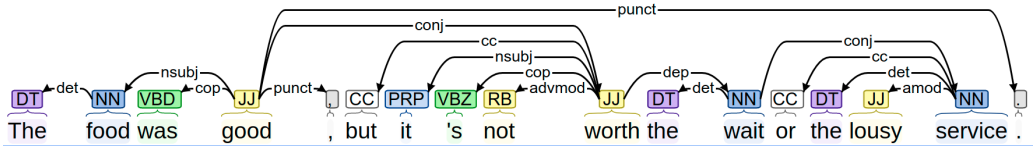

Figure 2: Sample Dependency Parser Output

by, Ganu et al. (2009) which contains ≈ 2.8 million unlabelled restaurant reviews.

4.2 Word Representations

We consider a pre-trained language model called BERT (Devlin et al., 2019) (in particular we chose the '*bert-base-uncased*' model which has 110M parameters). BERT follows a transformer model (Vaswani et al., 2017) for its representation, where the model predicts the masked words using the surrounding context words. We obtain vector representation for each word of a given sentence using BERT language model. We use BERT for both word representations and the post-training tasks.

4.3 Baselines

We compare the performance of the proposed model with diverse types of baselines such as random, supervised, and weakly supervised methods.

- **Random:** At first, we present a random baseline where the predictions are generated using a uniform distribution. This will provide us with a lower bound for our evaluation.
- **Supervised:** A recent supervised method, **ACSA-generation** (Liu et al., 2021) solves the ACSA as a generation task. The training and test set are structured with some predefined templates. Finally, the authors used BART (Lewis et al., 2020), a denoising autoencoder, for generating the desired outputs. This will give us an approximate upper bound for our evaluation.
- **Weakly Supervised:** A weakly supervised method, **JASen** (Huang et al., 2020) takes unlabelled training reviews and a few keywords corresponding to each aspect categories and sentiment polarity and outputs an (aspect, sentiment) pair for each review. The authors only considered the sentences with single aspect category.
- **Extremely Weakly Supervised:** The method **X-Class** (Wang et al., 2021b) takes reviews and a single keyword per class name as inputs and predicts a single class for each review. The method was validated majorly on different topic modelling datasets.

5 Experimental Evaluation

In this section, we study the performance of the different algorithms on four datasets, compare them with different baselines, and discuss the qualitative analysis of our model performance.

5.1 Evaluation Framework

We evaluate our method in an **End-to-End** framework. The popularly used ABSA evaluation uses gold aspects as a part of input to predict the sentiment polarity of each gold aspect. However, when the task is unsupervised (almost), we do not expect to know the aspect categories beforehand, as has been explored in previous works involving sentiment mining alone. Thus, we follow the End-to-End framework, which has two stages. In the first stage, given sentences, all the aspects are predicted. In the second stage, for each predicted aspect in the first stage, the corresponding sentiment polarity is predicted. Therefore, in our case, the first-stage output is the ACD output, which outputs aspect categories corresponding to each sentence. The second stage output is the ACSA output, which is a set of tuples consisting of (aspect category, sentiment polarity) pairs for each sentence. Therefore, if both the aspect category and sentiment polarity are predicted correctly then only, we consider it as a correct prediction. Thus, the performance is measured over all tuples (aspect, sentiment) in the gold data.

5.2 Evaluation Metrics

We consider two metrics for performance evaluation. For the ACD task, we report macro-averaged F1 score (or F1-macro) which is the average of F1-scores per class. For the ACSA task, we report macro-averaged F1-PN score (or macro F1-PN)

	Supervision Type	Methods	ACD				ACSA			
			Rest-14	Rest-15	Rest-16	MAMS	Rest-14	Rest-15	Rest-16	MAMS
Baselines	Random		22.50	21.12	19.03	16.45	08.40	08.46	07.16	05.39
	Supervised	ACSA-Generation	91.41	83.56	87.11	89.23	78.43	71.91	73.76	70.30
	Weakly Supervised	JASen	42.27	33.29	43.43	21.57	26.62	19.44	23.23	14.74
	Extremely Weakly Supervised	X-Class	46.69	40.35	36.58	36.52	34.44	25.49	24.83	16.32
Proposed	Extremely Weakly Supervised	X-SABSA	56.16	58.87	42.77	37.72	39.66	42.55	31.46	19.60
		AX-SABSA	69.57	56.17	45.69	39.33	44.14	40.24	32.23	18.55
		X-MABSA	61.73	**62.07**	49.02	56.48	44.96	**44.35**	35.81	27.28
		AX-MABSA	**74.90**	60.08	**50.63**	**60.82**	**49.68**	42.74	**36.47**	**29.74**

Table 2: Comparative Results for the ACD and End-to-End ACSA tasks. We report F1-macro score for ACD and F1-PN macro score for ACSA. X-SABSA: Proposed single label predictor model. AX-SABSA: Proposed single label predictor model where the candidate word for each class is also updated. X-MABSA: Proposed multi-label predictor model. AX-MABSA: Proposed multi-label predictor model where the candidate word for each class is also updated. Clustering algorithm used: mini batch k-means for ACD, and gmm for ACSA.

which is the mean of F1-scores of all aspect category, sentiment (positive, negative) pair tuples. The macro F1-PN is commonly used in different SemEval tasks (Pontiki et al., 2016).

5.3 Empirical Results

Comparative results of the ACD and ACSA tasks on different datasets are presented in Table 2. The results show that we achieve far better performance than random baselines given that our approach is unsupervised. The improvement of our multi-label models (X-MABSA and AX-MABSA) is statistically significant at $p < 0.01$ using paired t-test (Hsu and Lachenbruch, 2014) compared to proposed single label models (X-SABSA and AX-SABSA) and weakly supervised baselines (X-Class, and JASen).

For the ACD task, we achieve baseline results for all the datasets (ACSA module). We obtain F1-macro of 46.69, 40.35, 36.58, and 36.52 on Rest-14, Rest-15, Rest-16, and MAMS dataset, respectively. The proposed X-SABSA model improves the performance significantly on all the datasets (F1-macro of 56.16, 58.87, 42.77, and 37.72 on Rest-14, Rest-15, Rest-16, and MAMS data, respectively). Within our proposed models, we find that our multi-label model X-MABSA performs better than single-label model X-SABSA on all datasets. Especially, on the MAMS dataset, it improves the performance significantly (F1-macro of 56.48). We also observe that the AX-MABSA model (i.e., when automatically selected candidate words are considered for class representation) further improves performance on Rest-14, Rest-16, and MAMS datasets (F1-macro of 74.90, 50.63, and 60.82). It shows that the AX-MABSA model is more generalized and works very well when class names are not present in the input data.

As the ACSA task is framed as an End-to-End pipeline, we expect the performance to be lower than the often-used ACSA evaluation procedure. We achieve the baseline results (ACSA module) which are F1-PN-macro scores of 34.44, 25.49, 24.83, and 16.32 on Rest-14, Rest-15, Rest-16, and MAMS, respectively. We find that the proposed X-SABSA model improves the performance significantly over the baseline (F1-PN-macro of 39.66, 42.55, and 31.46 on Rest-14, Rest-15, and Rest-16, respectively). The multi-label model, X-MABSA improves the results further (F1-PN-macro of 44.96, 44.35, 35.81, and 27.28, respectively). We also observe that the AX-MABSA model improves the performance on Rest-14, Rest-16, and MAMS data (F1-PN-macro of 49.68, 36.47, and 29.74, respectively).

We observe that our proposed model performs significantly better than the random, and two weakly supervised baselines (X-Class and JASen) on both ACD and ACSA tasks. As our method is an extremely weakly supervised method, we do not expect our model to be better than the supervised model. However, in comparison to the supervised model (ACSA-generation), our method shows promising performance. For example, on the Rest-14 data, the supervised model achieves an F1-macro of 91.42 while our proposed model achieves an F1-macro of 74.90 for the ACD task. For the ACSA task, the proposed method performs decently compared to the supervised baseline. For example, on Rest-15, the supervised method achieves an F1-PN-macro of 71.91 while our method achieves an F1-PN-macro of 44.35.

It is evident that our proposed method works comparatively poorly for ACSA task on the MAMS data. The reason for this is the presence of a remarkably high number of 'neutral' classes (43.62% of total polarity labels). Selecting a single repre-

Review	Actual	Predicted
The sashimi is always fresh and the rolls are innovative and delicious.	(food, positive)	(food, positive)
While there's a decent menu, it shouldn't take ten minutes to get your drinks and 45 for a dessert pizza.	(food, positive), (service, negative)	(food, positive), (service, positive)
Who can't decide on a single dish, the tapas menu allowed me to express my true culinary self.	(food, negative), (menu, positive)	(menu, negative)
Roof: very nice space (although I know 5 other rooftop bars just as good), but the crowd was bunch of posers and the owner was a tool.	(place, positive), (miscellaneous, neutral)	(place, positive), (ambience, negative)
Endless fun, awesome music, great staff!	(service, positive), (ambience, positive), (restaurant, positive)	(service, positive), (ambience, positive)

Table 3: Illustration of the proposed method using few examples

sentative surface word for 'neutral' class is difficult as there is no association between any word and neutral sentences as compared to the 'positive' and 'negative' class. For example, the word 'bad' can be a representative of 'negative' class and the word 'good' can be the same of 'positive' class, but we found no such representative word for neutral class to perform well.

5.4 Performance Analysis

We report a few example texts with original and our model predicted tags in Table 3. We find that in some cases our model combines two closely related categories to one. For example, the text "*who can't decide on a single dish, the tapas menu allowed me to express my true culinary self.*" has gold category as *food* and *menu*. Our model predicts it as *menu*. The reason is that both the words 'dish', and 'menu' got higher similarity score to category 'menu' which is reasonable.

The fourth sentence in the Table 3 has gold labels as '(place, positive)' and '(miscellaneous, neutral)'. Our model predicts as '(place, positive)' and '(ambience, negative)'. We see here that the 'miscellaneous' class has been miss-classified into 'ambience' and 'neutral' to 'negative'. The 'miscellaneous' class is difficult to represent, even if we replace this word by the automatic surface word selection algorithm. Also, from the sentence, we can sense that the 'ambience' can be a class with 'negative' polarity.

The fifth sentence in Table 3 has gold labels as 'service', 'ambience' and 'restaurant'. However, our model predicts it as 'service', and 'ambience' missing the 'restaurant' category. This happened as in the sentence, there is no explicit presence of restaurant related words. Another point is that there are some mutual words related to both 'ambience' and 'restaurant', such as the word 'place' can be related to both 'ambience' and 'restaurant'.

6 Conclusion

In this paper, we studied extremely weakly supervised aspect category sentiment analysis across four benchmark datasets, and presented the state-of-the-art unsupervised framework without the requirement of any labelled data. Our method relied only on the surface text of aspect class names and unlabelled texts to extract aspect-sentiment pairs via a multi-label generator model. We proposed an automatic class-representative surface word selection algorithm to select proper representative words corresponding to each class. We also found that unsupervised post-training of language models on domain-specific data improved the word-representations and thus improved the performance. Experiments show that our proposed method performs better than all weakly supervised baseline models. In the future, we intend to improve our methods to incorporate more sentiment classes. We believe that our work would foster more research interest towards unsupervised ABSA.

7 Limitations

The main limitation of the proposed work is that it is unable to model the "neutral" sentiment class, and performs significantly lower when the number of neutral sentiment reviews are high in a dataset. This is evident from ACSA results on the MAMS data (in Table 2), where the number of neutral classes is high. We have also tried with some possible neutral class related seed category words like 'okay', 'moderate', 'average', etc. but the performance did not improve. It shows that these words can not represent the 'neutral' class. Thus, modelling the 'neutral' class efficiently will improve the model performance. Although our model performs better than other weakly supervised baselines, there is enough scope for improvement to bridge the gap between the supervised methodologies.

References

Hervé Abdi and Lynne J Williams. 2010. Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, 2(4):433–459.

Hongjie Cai, Yaofeng Tu, Xiangsheng Zhou, Jianfei Yu, and Rui Xia. 2020. Aspect-category based sentiment analysis with hierarchical graph convolutional network. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 833–843.

Hongjie Cai, Rui Xia, and Jianfei Yu. 2021. Aspect-category-opinion-sentiment quadruple extraction with implicit aspects and opinions. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 340–350.

Peng Chen, Zhongqian Sun, Lidong Bing, and Wei Yang. 2017. Recurrent attention network on memory for aspect sentiment analysis. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 452–461.

Zhuang Chen and Tieyun Qian. 2019. Transfer capsule network for aspect level sentiment classification. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 547–556.

Zhuang Chen and Tieyun Qian. 2020a. Enhancing aspect term extraction with soft prototypes. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2107–2117.

Zhuang Chen and Tieyun Qian. 2020b. Relation-aware collaborative learning for unified aspect-based sentiment analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3685–3694.

Jiajun Cheng, Shenglin Zhao, Jiani Zhang, Irwin King, Xin Zhang, and Hui Wang. 2017. Aspect-level sentiment classification with heat (hierarchical attention) network. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 97–106.

Hongliang Dai and Yangqiu Song. 2019. Neural aspect and opinion term extraction with mined rules as weak supervision. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5268–5277.

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)*, pages 4171–4186.

Richard O Duda, Peter E Hart, and David G Stork. 1973. *Pattern classification and scene analysis*, volume 3. Wiley New York.

Zhifang Fan, Zhen Wu, Xinyu Dai, Shujian Huang, and Jiajun Chen. 2019. Target-oriented opinion words extraction with target-fused neural sequence labeling. 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)*, pages 2509–2518.

Gayatree Ganu, Noemie Elhadad, and Amélie Marian. 2009. Beyond the stars: improving rating predictions using review text content. In *WebDB*, volume 9, pages 1–6. Citeseer.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910.

Ruidan He, Wee Sun Lee, Hwee Tou Ng, and Daniel Dahlmeier. 2018. Exploiting document knowledge for aspect-level sentiment classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 579–585.

Ruidan He, Wee Sun Lee, Hwee Tou Ng, and Daniel Dahlmeier. 2019. An interactive multi-task learning network for end-to-end aspect-based sentiment analysis. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 504–515.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Xiaochen Hou, Peng Qi, Guangtao Wang, Rex Ying, Jing Huang, Xiaodong He, and Bowen Zhou. 2021. Graph ensemble learning over multiple dependency trees for aspect-level sentiment classification. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2884–2894.

Henry Hsu and Peter A Lachenbruch. 2014. Paired t test. *Wiley StatsRef: statistics reference online*.

Jiaxin Huang, Yu Meng, Fang Guo, Heng Ji, and Jiawei Han. 2020. Weakly-supervised aspect-based sentiment analysis via joint aspect-sentiment topic embedding. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6989–6999.

Yann LeCun, Yoshua Bengio, et al. 1995. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 3361(10):1995.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

Kun Li, Chengbo Chen, Xiaojun Quan, Qing Ling, and Yan Song. 2020a. Conditional augmentation for aspect term extraction via masked sequence-to-sequence generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7056–7066.

Xin Li, Lidong Bing, Piji Li, Wai Lam, and Zhimou Yang. 2018. Aspect term extraction with history attention and selective transformation. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 4194–4200.

Xin Li and Wai Lam. 2017. Deep multi-task learning for aspect term extraction with memory interaction. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 2886–2892.

Yuncong Li, Cunxiang Yin, Sheng-hua Zhong, and Xu Pan. 2020b. Multi-instance multi-label learning networks for aspect-category sentiment analysis. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3550–3560.

Bin Liang, Hang Su, Rongdi Yin, Lin Gui, Min Yang, Qin Zhao, Xiaoqi Yu, and Ruifeng Xu. 2021. Beta distribution guided aspect-aware graph for aspect category sentiment analysis with affective knowledge. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 208–218.

Yunlong Liang, Fandong Meng, Jinchao Zhang, Jinan Xu, Yufeng Chen, and Jie Zhou. 2019. A novel aspect-guided deep transition model for aspect based sentiment analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5569–5580.

Bing Liu. 2012. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167.

Jian Liu, Zhiyang Teng, Leyang Cui, Hanmeng Liu, and Yue Zhang. 2021. Solving aspect category sentiment analysis as a text generation task. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4406–4416.

Jiangming Liu and Yue Zhang. 2017. Attention modeling for targeted sentiment. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 572–577.

Stuart Lloyd. 1982. Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137.

Ling Luo, Xiang Ao, Yan Song, Jinyao Li, Xiaopeng Yang, Qing He, and Dong Yu. 2019. Unsupervised neural aspect extraction with sememes. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 5123–5129.

Dehong Ma, Sujian Li, Fangzhao Wu, Xing Xie, and Houfeng Wang. 2019. Exploring sequence-to-sequence learning in aspect term extraction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3538–3547.

Dehong Ma, Sujian Li, Xiaodong Zhang, and Houfeng Wang. 2017. Interactive attention networks for aspect-level sentiment classification. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, pages 4068–4074.

Yukun Ma, Haiyun Peng, and Erik Cambria. 2018. Targeted aspect-based sentiment analysis via embedding commonsense knowledge into an attentive lstm. In *Proceedings of the AAAI conference on artificial intelligence*, volume 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)*, pages 19–30, San Diego, California. Association for Computational Linguistics.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Suresh Manandhar, and Ion Androutsopoulos. 2015. SemEval-2015 task 12: Aspect based sentiment analysis. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 486–495, Denver, Colorado. Association for Computational Linguistics.

Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. SemEval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 27–35, Dublin, Ireland. Association for Computational Linguistics.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human

languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

David Sculley. 2010. Web-scale k-means clustering. In *Proceedings of the 19th international conference on World wide web*, pages 1177–1178.

Tian Shi, Liuqing Li, Ping Wang, and Chandan K Reddy. 2021. A simple and effective self-supervised contrastive learning framework for aspect detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13815–13824.

Chi Sun, Luyao Huang, and Xipeng Qiu. 2019. Utilizing bert for aspect-based sentiment analysis via constructing auxiliary sentence. 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)*, pages 380–385.

Yi Tay, Luu Anh Tuan, and Siu Cheung Hui. 2018. Learning to attend via word-aspect associative fusion for aspect-based sentiment analysis. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Kai Wang, Weizhou Shen, Yunyi Yang, Xiaojun Quan, and Rui Wang. 2020. Relational graph attention network for aspect-based sentiment analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3229–3238.

Qianlong Wang, Zhiyuan Wen, Qin Zhao, Min Yang, and Ruifeng Xu. 2021a. Progressive self-training with discriminator for aspect term extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 257–268.

Yequan Wang, Minlie Huang, Xiaoyan Zhu, and Li Zhao. 2016. Attention-based lstm for aspect-level sentiment classification. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 606–615.

Zihan Wang, Dheeraj Mekala, and Jingbo Shang. 2021b. X-class: Text classification with extremely weak supervision. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3043–3053.

Zhen Wu, Chengcan Ying, Fei Zhao, Zhifang Fan, Xinyu Dai, and Rui Xia. 2020a. Grid tagging scheme for aspect-oriented fine-grained opinion extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2576–2585.

Zhen Wu, Fei Zhao, Xin-Yu Dai, Shujian Huang, and Jiajun Chen. 2020b. Latent opinions transfer network for target-oriented opinion words extraction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9298–9305.

Hu Xu, Bing Liu, Lei Shu, and S Yu Philip. 2018. Double embeddings and cnn-based sequence labeling for aspect extraction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 592–598.

Lu Xu, Yew Ken Chia, and Lidong Bing. 2021. Learning span-level interactions for aspect sentiment triplet extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4755–4766.

Wei Xue and Tao Li. 2018. Aspect based sentiment analysis with gated convolutional networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2514–2523.

Hang Yan, Junqi Dai, Tuo Ji, Xipeng Qiu, and Zheng Zhang. 2021. A unified generative framework for aspect-based sentiment analysis. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2416–2429.

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

He Zhao, Longtao Huang, Rong Zhang, Quan Lu, and Hui Xue. 2020. Spanmlt: A span-based multi-task learning framework for pair-wise aspect and opinion terms extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3239–3248.