AspectCSE: Sentence Embeddings for Aspect-based Semantic Textual Similarity Using Contrastive Learning and Structured Knowledge

Tim Schopf[1], **Emanuel Gerber**[1], **Malte Ostendorff**[2], **and Florian Matthes**[1]
[1]Technical University of Munich, Department of Computer Science, Garching, Germany
[2]DFKI GmbH, Berlin, Germany
{tim.schopf,emanuel.gerber,matthes}@tum.de
malte.ostendorff@dfki.de

Abstract

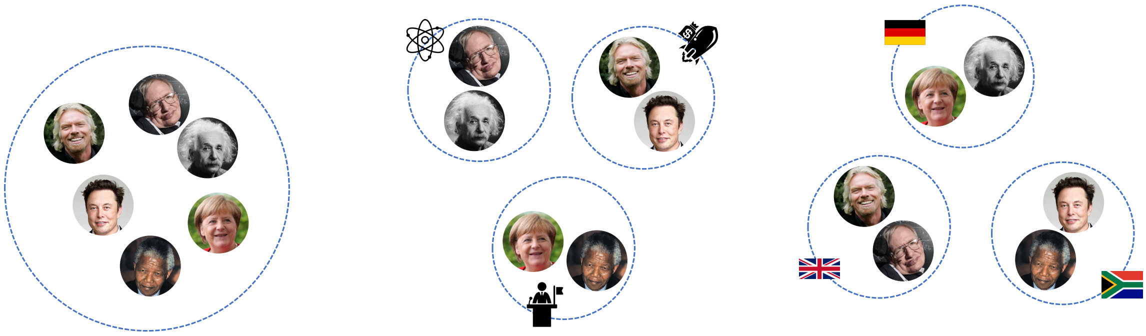
Generic sentence embeddings provide a coarse-grained approximation of semantic textual similarity but ignore specific aspects that make texts similar. Conversely, aspect-based sentence embeddings provide similarities between texts based on certain predefined aspects. Thus, similarity predictions of texts are more targeted to specific requirements and more easily explainable. In this paper, we present AspectCSE, an approach for aspect-based contrastive learning of sentence embeddings. Results indicate that AspectCSE achieves an average improvement of 3.97% on information retrieval tasks across multiple aspects compared to the previous best results. We also propose using Wikidata knowledge graph properties to train models of multi-aspect sentence embeddings in which multiple specific aspects are simultaneously considered during similarity predictions. We demonstrate that multi-aspect embeddings outperform single-aspect embeddings on aspect-specific information retrieval tasks. Finally, we examine the aspect-based sentence embedding space and demonstrate that embeddings of semantically similar aspect labels are often close, even without explicit similarity training between different aspect labels.

1 Introduction

Sentence embeddings are representations of sentences or short text paragraphs in a dense vector space, such that similar sentences are close to each other (Reimers and Gurevych, 2020). Learning sentence embeddings is a fundamental task in natural language processing (NLP) and has already been extensively investigated in the literature (Kiros et al., 2015; Hill et al., 2016; Conneau et al., 2017; Logeswaran and Lee, 2018; Cer et al., 2018; Reimers and Gurevych, 2019; Gao et al., 2021; Schopf et al., 2023d). Generic sentence embeddings can be used to distinguish between similar and dissimilar sentences, without considering which aspects of sentences are similar (Ostendorff et al., 2020a). Moreover, they are often evaluated on generic semantic textual similarity (STS) tasks (Marelli et al., 2014; Agirre et al., 2012, 2013, 2014, 2015, 2016; Cer et al., 2017) in which sentence similarity scores rely on human annotations. However, the concept of generic STS is not well defined, and text similarity depends heavily on the aspects that make them similar (Bär et al., 2011; Ostendorff et al., 2020b, 2022). We follow the argument of Bär et al. (2011) on textual similarity and define *aspects* as inherent properties of texts that must be considered when predicting their semantic similarity. Based on the different aspects focused on in texts, their similarities can be perceived very differently. Figure 1 illustrates an example of aspect-based STS. For example, Wikipedia introduction texts of famous individuals can generally be considered similar as all texts introduce people who are known to the public. However, focusing the comparison on specific aspects (e.g., *country of birth* or *profession*) leads to different semantic similarity assessments for the same texts. Although Wikipedia is a special case as the introduction texts represent specific entities, this characteristic can nevertheless be generalized to different aspects found in any text. When deciding the similarity of texts, different aspects must be considered. Consequently, human-annotated STS datasets introduce considerable subjectivity regarding the evaluated aspects.

Prior work uses siamese networks and a multiple negative ranking loss (Henderson et al., 2017) with only positive samples from the train set to create sentence embeddings for single aspects (Ostendorff et al., 2022). Sentence embeddings for single aspects only consider one specific aspect during similarity comparisons. Using structured knowledge from knowledge graphs (KGs) for language model training has been shown to improve performances on all types of downstream tasks (Schnei-

Proceedings of Recent Advances in Natural Language Processing, pages 1054–1065
Varna, Sep 4–6, 2023
https://doi.org/10.26615/978-954-452-092-2_113



(a) Generic sentence embeddings (b) Sentence embeddings based on the *profession* aspect. (c) Sentence embeddings based on the *country of birth* aspect.

Figure 1: Images of famous people with the corresponding Wikipedia introductory texts as sentence embeddings in a dense vector space. Blue dashed circles represent clusters of semantically similar embeddings. Based on the encoded aspect, embeddings of these same texts can be distributed differently in a vector space. (a) All generic embeddings are close and approximately evenly distributed as the texts introduce famous people. (b) Embeddings that focus on the *profession* aspect are close if the people have similar professions. (c) Embeddings that focus on the *country of birth* aspect are close if the people have similar countries of birth.

der et al., 2022) and also provides the possibility to create sentence embeddings that focus on multiple specific aspects simultaneously. These sentence embeddings are especially useful in information retrieval or unsupervised text classification settings (Schopf et al., 2021, 2022, 2023a,b,c).

In this work, we advance state-of-the-art sentence embeddings for aspect-based STS using AspectCSE, an approach for aspect-based contrastive learning of sentence embeddings. Additionally, we introduce multi-aspect sentence embeddings that simultaneously consider multiple specific aspects during similarity comparisons. We show the effectiveness of multi-aspect sentence embeddings for both information retrieval and exploratory search tasks. Finally, we demonstrate that using KG properties can be extremely beneficial for creating both single- and multi-aspect sentence embeddings.

2 Related Work

In NLP, *aspects* are most commonly examined in sentiment analysis problems (Pontiki et al., 2014; Xue and Li, 2018; Brun and Nikoulina, 2018; Zhang et al., 2021; Yan et al., 2021; Liang et al., 2022). Thus, the goal is to identify the aspects of given target entities and the sentiment expressed for each aspect (Pontiki et al., 2014).

Some works investigate aspect-based STS by considering it as a segmentation task. Chan et al. (2018) first segmented abstracts of research papers according to different aspects. Then, they constructed semantic representations from these aspect-based segments, which can be used to find analogies between research papers. Huang et al. (2020) presented a human-annotated dataset that segments 10,966 English abstracts in the COVID-19 Open Research Dataset (Wang et al., 2020) by the aspects background, purpose, method, result/contribution, and others. Kobayashi et al. (2018) learned multi-vector representations of segmented scientific articles in which each vector encodes a different aspect. However, segmenting texts can harm their coherence and decrease the performance of downstream NLP models (Gong et al., 2020).

Other approaches propose to treat aspect-based STS as a pairwise multi-class classification problem (Ostendorff et al., 2020a,b). However, Reimers and Gurevych (2019) argue that pairwise classification with transformer models results in quadratic complexity. Therefore, this approach is not suitable for large-scale STS tasks.

To address the issues using previous approaches, Ostendorff et al. (2022) proposed training aspect-based embeddings for research papers. In this work, we use AspectCSE and KG properties to train single- and multi-aspect sentence embeddings. This allows us to focus on multiple specific aspects simultaneously while improving the performance of aspect-based sentence embeddings in STS tasks.

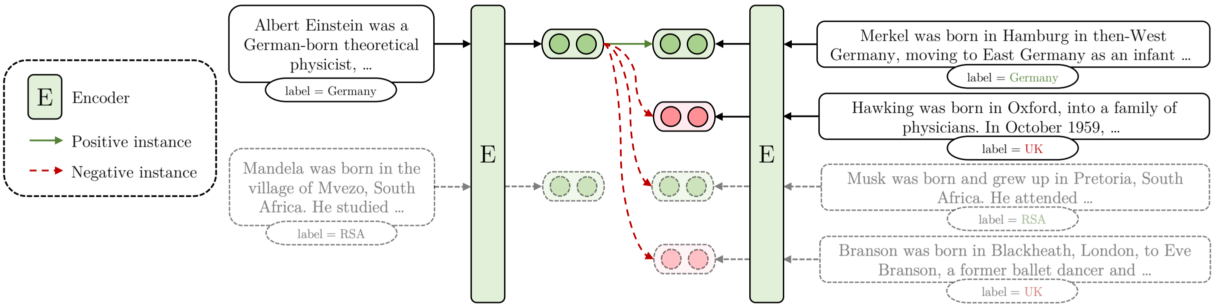


Figure 2: AspectCSE uses (*anchor, positive, negative*) triplets to train aspect-specific sentence embedding models. Pairs with the same label for a specific aspect (here: *country of birth*) are used as positives and those with different labels for the same aspect and other in-batch instances as negatives.

3 Embedding Methods

3.1 AspectCSE

Recently, contrastive learning has exhibited state-of-the-art performance for generic sentence embeddings (Gao et al., 2021; Giorgi et al., 2021; Chuang et al., 2022). The contrastive learning objective creates effective representations by pulling semantically close neighbors together and pushing apart non-neighbors (Hadsell et al., 2006). We follow the proposed supervised contrastive learning framework of Gao et al. (2021) and use a cross-entropy-loss with negatives per anchor-positive pair and random in-batch negatives. To train aspect-based sentence embedding models, we assume a set of triplets $\mathcal{D} = \{(x_i^a, x_i^{a+}, x_i^{a-})\}$. Here, x_i^a is an anchor sentence, x_i^{a+} is semantically related, and x_i^{a-} is semantically unrelated to x_i^a with respect to aspect a. With $\mathbf{h}_i^a$, $\mathbf{h}_i^{a+}$, and $\mathbf{h}_i^{a-}$ as representations of x_i^a,x_i^{a+}, and x_i^{a-}, the training objective with a mini-batch of N triplets is expressed as:

$$\ell_i = -\log \frac{e^{sim(\mathbf{h}_i^a,\mathbf{h}_i^{a+})/\tau}}{\sum_{j=1}^{N}(e^{sim(\mathbf{h}_i^a,\mathbf{h}_j^{a+})/\tau} + e^{sim(\mathbf{h}_i^a,\mathbf{h}_j^{a-})/\tau})} \quad (1)$$

where τ is a temperature hyperparameter and $sim(\mathbf{h}_1, \mathbf{h}_2)$ is the cosine similarity $\frac{\mathbf{h}_1 \cdot \mathbf{h}_2}{||\mathbf{h}_1|| \cdot ||\mathbf{h}_2||}$. To encode input sentences, we use BERT-based pretrained language models (Devlin et al., 2019) and fine-tune the parameters using the contrastive objective (Equation 1). Figure 2 illustrates the proposed AspectCSE approach.

3.2 Multiple Negative Ranking Using Anchor-Positive Pairs Only

As a baseline, we perform aspect-based fine-tuning of BERT-based pretrained language models following the state-of-the-art approach of Ostendorff et al. (2022). Therefore, we use mean pooling and a multiple negative ranking loss (Henderson et al., 2017) with anchor-positive pairs for training. Therefore, the training input comprises a set of positive samples $\mathcal{D} = \{(x_i^a, x_i^{a+})\}$ only. During training, every instance $x_j^{a+} = \{x_1^{a+} ... x_{N-1}^{a+}\}$ within a mini-batch of N samples is used as random negative for anchor x_i^a if $i \neq j$.

4 Data

For our experiments, we use two different datasets. First, we use a benchmark dataset derived from Papers with Code (PwC) [1] to evaluate the effectiveness of AspectCSE. We also use Wikipedia and the Wikidata KG (Vrandečić and Krötzsch, 2014) to build a dataset for learning multi-aspect sentence embeddings. In all our experiments, we consider a pair of texts as positive if they share the same label for a particular aspect. Accordingly, negatives comprise a pair of texts with different labels for a particular aspect.

4.1 Papers with Code

The PwC dataset is a collection of research paper abstracts that are annotated with *task*, *method* and *dataset* aspects and their respective labels (Ostendorff et al., 2022). In this dataset, for example, a label of the *task* aspect is *self-supervised learning* or *machine translation*. We obtain the dataset version from 2022-05-25 and remove paper abstracts that belong to aspect labels with more than 100 instances. Abstracts with less than 100 characters are also removed. Table 1 summarizes the resulting PwC dataset. We split the final PwC dataset into 80% training and 20% test paper abstracts for our experiments.

[1] https://paperswithcode.com

Aspect	# Papers	# Labels
Task	32,873	2,481
Method	10,213	1,724
Dataset	7,305	3,611

Table 1: Summary of the PwC dataset.

4.2 Wikipedia and Wikidata

Wikipedia contains a broad range of topics with many possible aspects for each article. We have found that the number of articles regarding companies in Wikipedia accounts for a large portion of the articles, while the introductory sections contain a reasonable amount of different aspects. Therefore, in our experiments focus on a subset of Wikipedia, which includes the introduction section of articles about companies only. Furthermore, we use the commonly occurring aspects *industry (e.g., What type of product/service does the company offer?)* and *country (e.g., What country is the company based in?)* for our experiments. Since Wikipedia comprises unstructured texts only, we take advantage of most Wikidata KG entities being linked to their corresponding Wikipedia articles. We also consider specific Wikidata properties as aspects while using the values linked to a seed article by the specific properties as labels. In this case, we use the Wikidata properties *country* and *industry* as aspects while taking the values linked to the company articles by these properties as labels. Therefore, we follow the approach in Algorithm 1 to construct our dataset.

Algorithm 1 Construct aspect-based dataset

Require:
  $companies$ = list of all Wikidata entities e of type *business* (Q4830453)
  $companies_{annotated} \leftarrow \emptyset$
  procedure ANNOTATE($companies$)
    for e **in** $companies$ **do**
      if e_k has Wikipedia article w_k **then**
        s = introduction section of w_k
        s_c = *country* (P17) value(s) of e_k
        s_i = *industry* (P452) value(s) of e_k
        $companies_{annotated} += (s, s_c, s_i)$
    return $companies_{annotated}$

We use the Wikidata SPARQL API to find the companies as well as the country and industry values linked to them. We also use the Kensho Derived Wikimedia Dataset[2], which comprises preprocessed Wikipedia and Wikidata dumps from 2019-12-01, to obtain the Wikipedia introduction sections of the retrieved companies. Moreover, we utilize the Kensho Derived Wikimedia Dataset to sample 10,000 random articles from different topics without any aspect information. In addition to the company introduction sections, these random articles are used as further negatives during training. This ensures that the model learns to distinguish between different aspect labels and between different topics. Table 2 summarizes the resulting dataset. For example, the labels for the *country* aspect are USA or Germany. For our experiments, we split the final dataset into 80% training and 20% test data.

Aspect	# Articles	# Labels
Industry	6,082	97
Country	2,062	75
Random articles	10,000	-

Table 2: Summary of the Wikipedia + Wikidata dataset.

To train aspect-based sentence embeddings with AspectCSE, we further process the dataset to yield triplets as follows:

- **Single-aspect-specific (Country)**: $(x_i^a, x_i^{a+}, x_i^{a-}) \Rightarrow x_i^{a+}$ and x_i^{a-} are positive and negative samples w.r.t. the country aspect a.
- **Single-aspect-specific (Industry)**: $(x_i^b, x_i^{b+}, x_i^{b-}) \Rightarrow x_i^{b+}$ and x_i^{b-} are positive and negative samples w.r.t. the industry aspect b.
- **Multi-aspect-specific (Intersection)**: $(x_i^{a,b}, x_i^{a+\cap b+}, x_i^{a-\cap b-}) \Rightarrow x_i^{a+\cap b+}$ is a positive sample if it has **both** the same country aspect a **and** the same industry aspect b as the seed sentence.
- **Multi-aspect-specific (Union)**: $(x_i^{a,b}, x_i^{a+\cup b+}, x_i^{a-\cup b-}) \Rightarrow x_i^{a+\cup b+}$ is a positive sample if it has **either** the same country aspect a **or** the same industry aspect b as the seed sentence.

[2]https://www.kaggle.com/datasets/kenshoresearch/kensho-derived-wikimedia-data

Aspects →		Task			Method			Dataset		
Methods ↓		P	R	MRR	P	R	MRR	P	R	MRR
Generic	$\text{SciBERT}_{\texttt{base}}$	0.071	0.070	0.244	0.051	0.056	0.181	0.060	0.101	0.212
	$\text{DeCLUTR}_{\texttt{sci-base}}$	0.130	0.131	0.369	0.069	0.078	0.219	0.099	0.170	0.317
	SPECTER	0.248	0.247	0.521	0.104	0.117	0.277	0.183	0.311	0.464
Aspect-based	Multiple Negative Ranking	0.409	0.424	0.768	0.263	0.302	0.595	0.172	0.418	0.465
	$*$ AspectCSE	**0.416**	**0.431**	**0.776**	**0.268**	**0.312**	**0.606**	**0.186**	**0.461**	**0.507**

Table 3: Evaluation results for retrieving the $k = 10$ most similar elements for different sentence embedding approaches on the PwC test dataset. *AspectCSE* indicates the training approach explained in Section 3.1. *Multiple Negative Ranking* indicates the training approach explained in Section 3.2. Precision@k (P), Recall@k (R), and Mean Reciprocal Rank@k (MRR) are reported.

5 Experiments

5.1 Comparison with Baselines

To evaluate AspectCSE against state-of-the-art baselines, we use the PwC benchmark dataset described in Section 4.1 for model training and testing.

Generic Sentence Embeddings We evaluate AspectCSE against multiple generic sentence embedding models from the scholarly domain. These models are pretrained on scientific literature and produce domain-specific state-of-the-art sentence embeddings without leveraging any aspect information. We use SciBERT (Beltagy et al., 2019), SPECTER (Cohan et al., 2020), and DeCLUTR (Giorgi et al., 2021) in their `base`-versions as published by their authors without any fine-tuning on our corpus. For SciBERT, we use the concatenated outputs of the last four layers as embeddings.

Parameter	Value
Training epochs	3
Batch size	14
Learning rate	$5e-5$
Max sequence length	320
Pooler type	CLS
Temperature for softmax	0.05
Floating precision	16

Table 4: AspectCSE fine-tuning configuration.

Aspect-based Sentence Embeddings In addition to generic baselines, we train aspect-based sentence embedding models for each PwC aspect using SciBERT and the multiple negative ranking approach, as described in Section 3.2. To train AspectCSE, we use SciBERT as base model and the fine-tuning configuration presented in Table 4. For aspect-specific baseline training with multiple negative ranking, we use the same configuration, except that we follow the approach of Ostendorff et al. (2022), and apply MEAN pooling. For AspectCSE, we follow the argument of Gao et al. (2021), who found that different pooling methods do not matter much and use CLS.

5.2 Multi-aspect Sentence Embeddings

We use the Wikipedia + Wikidata dataset described in Section 4.2 to train and evaluate multi-aspect sentence embeddings. Further, we use AspectCSE to train multi- and single-aspect sentence embedding models for the *country* and *industry* aspects. For fine-tuning, we use $\text{BERT}_{\texttt{base}}$ and the training configuration presented in Table 4. To evaluate the performance of generic sentence embeddings on the Wikipedia + Wikidata test dataset, we use a trained $\text{SimCSE}_{\texttt{sup-bert-base}}$ model (Gao et al., 2021), which generates state-of-the-art generic sentence embeddings.

6 Evaluation

6.1 Information Retrieval Performance

For evaluation, we follow the approach of Ostendorff et al. (2022) and frame it as an information retrieval task. Therefore, we retrieve the $k = 10$ nearest neighbors for each element in the respective test datasets. After that, we determine the number of retrieved elements that match the particular aspect label of the seed element. We use the following evaluation metrics for this purpose:

- **Precision@k** (P): The number of nearest neighbors (within the top k candidates) that share the same aspect as the seed document divided by k.

Aspects →	*Country*			*Industry*		
Embedding type ↓	**P**	**R**	**MRR**	**P**	**R**	**MRR**
SimCSE$_{\texttt{generic}}$	0.315	0.058	0.523	0.320	0.061	0.531
AspectCSE$_{\texttt{single-aspect}}$	0.390	0.124	0.558	**0.625**	**0.178**	0.729
AspectCSE$_{\texttt{multi-aspect (Intersection)}}$	0.444	0.102	0.593	0.622	0.174	0.720
AspectCSE$_{\texttt{multi-aspect (Union)}}$	**0.555**	**0.163**	**0.738**	0.538	0.155	**0.747**

Table 5: Evaluation results for retrieving the $k = 10$ most similar elements for different sentence embedding approaches on the Wikipedia + Wikidata test dataset. Precision@k (P), Recall@k (R), and Mean Reciprocal Rank@k (MRR) are reported.

- **Recall@k** (R): The number of nearest neighbors (within the top k candidates) that share the same aspect as the seed document divided by the number of labeled documents with the seed document's aspect.
- **Mean Reciprocal Rank@k** (MRR): Measure of the ranking quality for the nearest neighbors, calculated by averaging the reciprocal ranks ($\frac{1}{\text{rank}}$) of each neighbor. This adds more weight to correctly labeled neighbors the higher they rank.

Papers with Code Table 3 compares AspectCSE, generic sentence embedding baselines, and the aspect-based multiple negative ranking baseline. The generic sentence embedding models perform badly for all evaluated aspects. Except for SPECTER, which achieves a respectable MRR score in the *dataset* aspect, generic models always perform significantly worse than aspect-based models. Therefore, aspect-based models retrieve similar texts of the same aspect much better than generic ones. Furthermore, By a large margin, AspectCSE outperforms the multiple negative ranking approach on all aspects and metrics. The average improvement is 3.97% for MRR scores of all PwC aspects. Hence, AspectCSE is a better approach for training aspect-based sentence embedding models. Accordingly, we use AspectCSE to train and evaluate multi-aspect sentence embedding models on the Wikipedia + Wikidata dataset.

Wikipedia and Wikidata Table 5 shows the evaluation results for the multi-aspect sentence embeddings on the Wikipedia + Wikidata test dataset. All AspectCSE models achieve strong performance in both aspects. While we train two separate embedding models for the single-aspect case (one embedding model each for the *country* and *industry* aspects), the multi-aspect models are trained on both

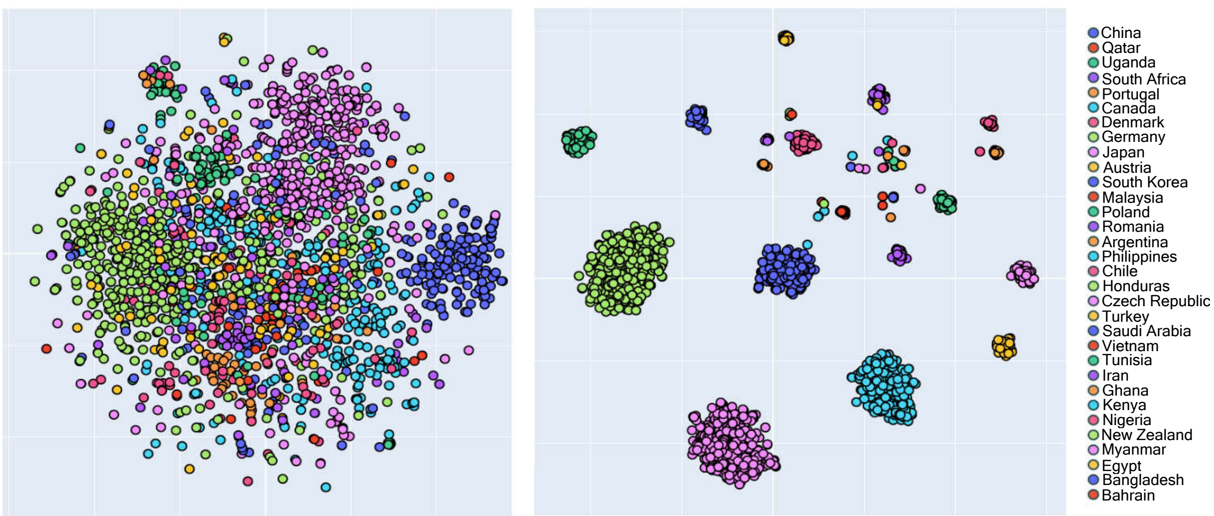


Figure 3: Comparison of generic sentence embeddings (left) vs. single-aspect sentence embeddings based on the *country* aspect (right).

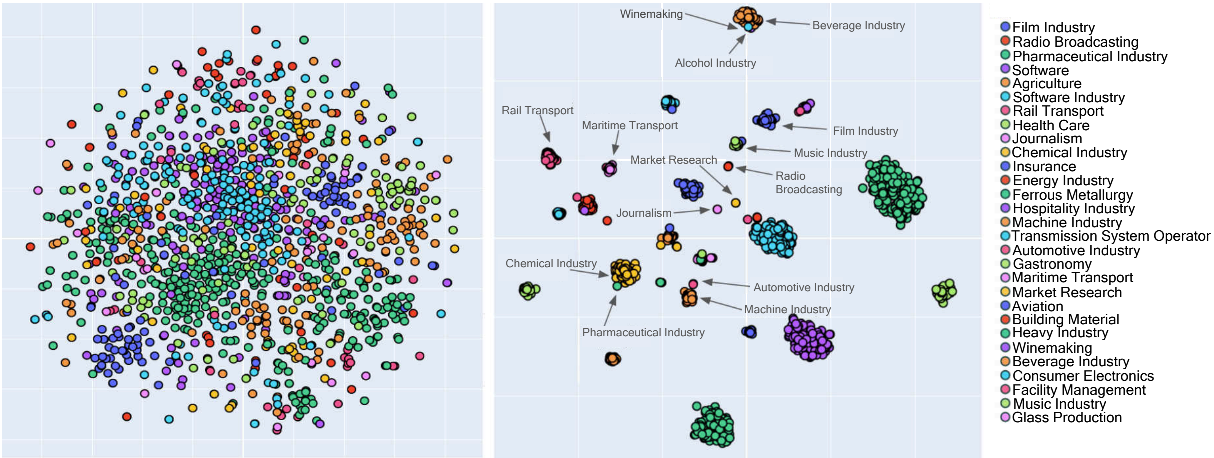


Figure 4: Comparison of generic sentence embeddings (left) vs. single-aspect sentence embeddings based on the *industry* aspect (right).

aspects simultaneously. Therefore, in the multi-aspect cases, only one model is used to retrieve the most similar elements for both aspects. Surprisingly, the best MRR scores for the *country* and *industry* aspects are achieved using the multi-aspect (Union) model, outperforming the multi-aspect (Intersection) and even the single-aspect models. A possible reason is that training sentence embedding models for multiple aspects provides the model with more training data. For example, a correlation exists between the type of industry and certain countries (e.g., Arab countries that have a higher than average density of oil companies) that may function as additional training data for the model.

6.2 Embedding Space Exploration

In addition to the information retrieval evaluation, we visually analyze selected generic, single-, and multi-aspect sentence embeddings. Therefore, we again use the Wikipedia + Wikidata dataset and the trained models described in Section 5.2. We utilize t-SNE (van der Maaten and Hinton, 2008) to reduce the dimensionality of sentence embeddings from 768 to 2 and color all data points according to their aspect labels. Figures 3 and 4 show the embedding spaces of generic and single-aspect sentence embeddings for the *country* and *industry* aspects. In these figures, generic sentence embeddings weakly capture both target aspects, as certain aspect labels dominate some regions. However, no clear separation can be observed between aspect labels and many aspect labels are scattered throughout the entire embedding space. Meanwhile, a sharp separation exists between aspect labels for aspect-based sentence embeddings with dense clusters of elements that share the same aspect label. This finding is consistent with our results in Table 5. Figure 4 shows the local neighborhoods of industry-specific sentence embeddings that reflect the semantic similarity of different industries. We observe that embeddings of the same aspect label are close to each other, and those of semantically similar aspect labels are closer when compared to embeddings with semantically dissimilar aspect labels. For example, embeddings with the semantically related aspect labels ”Film Industry”, ”Music Industry”, and ”Radio Broadcasting” are close to each other, whereas ”Rail Transport” and ”Maritime Transport” are located next to each other.

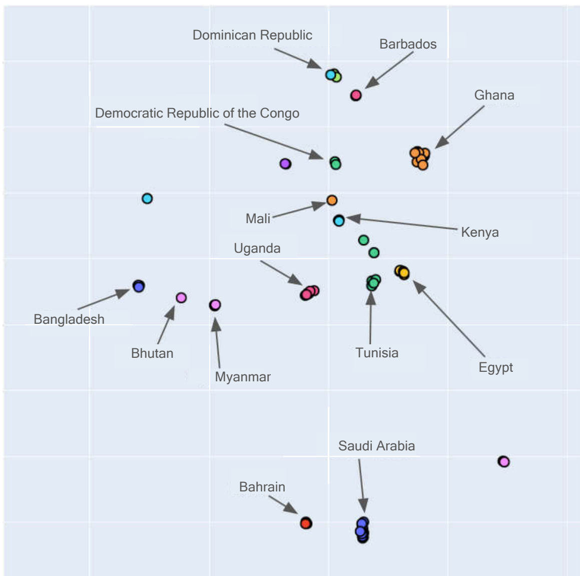


Figure 5: Local embedding space for single-aspect sentence embeddings based on the *country* aspect. The colors represent different aspect labels for the *country* aspect.

Figure 5 shows the local neighborhoods of single-aspect sentence embeddings based on the *country* aspect. We observe a similar behavior as in Figure 4, where embeddings of semantically similar aspect labels are close. For example, country-specific sentence embeddings of African countries (e.g., Kenya, Egypt, and Mali), Arab countries (e.g., Saudi Arabia, Bahrain), and South American countries (e.g., Dominican Republic, Barbados) share local neighborhoods, respectively. Although a correlation exists between semantically similar aspect labels and local neighborhoods in many cases, this pattern is not consistent for all aspect labels. For example, embeddings for the aspect label ”Austria” are closer to the embeddings from ”Japan” than to those for ”Germany”. This similarity pattern is likely a result of the fact that some texts from our training dataset are annotated with multiple aspects (e.g., ”Amazon” is annotated with ”e-commerce”, ”retail”, and ”cloud computing”). Since the model optimizes the embedding for Amazon to be close to e-commerce, retail, and cloud computing companies, all embeddings from these industries are pulled closer together. As the same company often operates in related industries (e.g., e-commerce and retail), this is likely why sentence embeddings of related aspect labels are close to each other. The pattern inconsistency may be partially a consequence of dimensionality reduction, where fine-grained differences between embeddings become lost.

Figure 6 shows the embeddings space for multi-aspect sentence embeddings (Union). This multi-aspect sentence embedding model (Union) learned to keep embeddings close to each other that share either the same *industry* or the same *country* or both aspects. As shown in the figure, only the *industry* aspect is colored, as it is the more dominant aspect for the spatial positioning of embeddings. Figure 6 shows the local neighborhoods that mostly contain embeddings of the same *industry* aspects. Simultaneously, the *country* aspect determines the spatial positioning of embeddings within the individual ”industry clusters”. Sentence embeddings that belong to a certain *industry* aspect, such as ”Automotive” are split into different country-specific sub-clusters. Furthermore, embeddings at the boundary between industries are likely to share the same *country* aspect. This is shown, for example, in ”Automotive Industry (China)” and ”Consumer Electronics (China)” embeddings located next to each other.

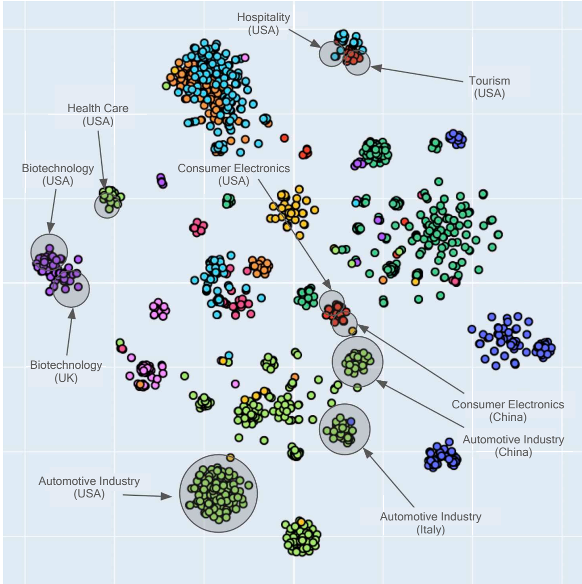


Figure 6: Global embedding space for multi-aspect sentence embeddings (Union). The colors represent different aspect labels for the *industry* aspect. The aspect-based sentence embedding model is trained with the contrastive learning approach stated in Section 3.1 and on the Wikipedia + Wikidata dataset described in Section 4.2

Overall, training AspectCSE using KG properties as aspects performs well in all our evaluations. Moreover, the multi-aspect (Union) model outperforms all other models by a large margin. Therefore, using KG properties and AspectCSE to train single-aspect and especially multi-aspect sentence embedding models achieves meaningful results in STS tasks.

7 Conclusion

In this work, we proposed using Wikidata knowledge graph properties to train single-aspect and multi-aspect sentence embedding models. Unlike single-aspect sentence embeddings, multi-aspect sentence embeddings consider multiple specific aspects simultaneously during similarity comparisons. We regarded STS as an information retrieval task and introduced the AspectCSE approach for training aspect-based sentence embedding models that achieve state-of-the-art performance on the PwC benchmark dataset. Furthermore, we demonstrated that training aspect-based sentence embedding models on multiple aspects simultaneously even surpasses the performance of single-aspect sentence embedding models. Finally, we show that the semantic similarity between different aspect labels is often connected to spatial proximity in the embedding space. This behavior is even clear if we train sentence embedding models only for similarity within the same aspect label but not explicitly for similarity between different aspect labels.

8 Limitations

AspectCSE only works for domains and languages with pretrained language models available for fine-tuning. Furthermore, using Wikidata KG properties to train single-aspect and multi-aspect sentence embedding models requires the availability of this structured information in large quantities. For widely used languages and domains, this requirement may be given. However, for under-represented languages and domains, Wikidata information is sparse, which has a negative impact on AspectCSE. Moreover, we evaluated our approach using texts that comprise entire paragraphs. Whether AspectCSE can also properly represent the specific aspects contained in individual sentences needs to be investigated in future work. Finally, training AspectCSE using CPU only is not feasible. Therefore, we used a Nvidia v100 GPU for AspectCSE training.

References

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Iñigo Lopez-Gazpio, Montse Maritxalar, Rada Mihalcea, German Rigau, Larraitz Uria, and Janyce Wiebe. 2015. SemEval-2015 task 2: Semantic textual similarity, English, Spanish and pilot on interpretability. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 252–263, Denver, Colorado. Association for Computational Linguistics.

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2014. SemEval-2014 task 10: Multilingual semantic textual similarity. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 81–91, Dublin, Ireland. Association for Computational Linguistics.

Eneko Agirre, Carmen Banea, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2016. SemEval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 497–511, San Diego, California. Association for Computational Linguistics.

Eneko Agirre, Daniel Cer, Mona Diab, and Aitor Gonzalez-Agirre. 2012. SemEval-2012 task 6: A pilot on semantic textual similarity. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 385–393, Montréal, Canada. Association for Computational Linguistics.

Eneko Agirre, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, and Weiwei Guo. 2013. *SEM 2013 shared task: Semantic textual similarity. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, pages 32–43, Atlanta, Georgia, USA. Association for Computational Linguistics.

Daniel Bär, Torsten Zesch, and Iryna Gurevych. 2011. A reflective view on text similarity. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, pages 515–520, Hissar, Bulgaria. Association for Computational Linguistics.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. 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 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Caroline Brun and Vassilina Nikoulina. 2018. Aspect based sentiment analysis into the wild. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 116–122, Brussels, Belgium. Association for Computational Linguistics.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada. Association for Computational Linguistics.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Brian Strope, and Ray Kurzweil. 2018. Universal sentence encoder for English. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 169–174, Brussels, Belgium. Association for Computational Linguistics.

Joel Chan, Joseph Chee Chang, Tom Hope, Dafna Shahaf, and Aniket Kittur. 2018. Solvent: A mixed initiative system for finding analogies between research papers. *Proc. ACM Hum.-Comput. Interact.*, 2(CSCW).

Yung-Sung Chuang, Rumen Dangovski, Hongyin Luo, Yang Zhang, Shiyu Chang, Marin Soljacic, Shang-Wen Li, Scott Yih, Yoon Kim, and James Glass. 2022. DiffCSE: Difference-based contrastive learning for sentence embeddings. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4207–4218, Seattle, United States. Association for Computational Linguistics.

Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel Weld. 2020. SPECTER: Document-level representation learning using citation-informed transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2270–2282, Online. Association for Computational Linguistics.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680, Copenhagen, Denmark. Association for Computational Linguistics.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

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, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

John Giorgi, Osvald Nitski, Bo Wang, and Gary Bader. 2021. DeCLUTR: Deep contrastive learning for unsupervised textual representations. 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 879–895, Online. Association for Computational Linguistics.

Hongyu Gong, Yelong Shen, Dian Yu, Jianshu Chen, and Dong Yu. 2020. Recurrent chunking mechanisms for long-text machine reading comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6751–6761, Online. Association for Computational Linguistics.

R. Hadsell, S. Chopra, and Y. LeCun. 2006. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742.

Matthew Henderson, Rami Al-Rfou, Brian Strope, Yun-hsuan Sung, Laszlo Lukacs, Ruiqi Guo, Sanjiv Kumar, Balint Miklos, and Ray Kurzweil. 2017. Efficient natural language response suggestion for smart reply.

Felix Hill, Kyunghyun Cho, and Anna Korhonen. 2016. Learning distributed representations of sentences from unlabelled data. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1367–1377, San Diego, California. Association for Computational Linguistics.

Ting-Hao Kenneth Huang, Chieh-Yang Huang, Chien-Kuang Cornelia Ding, Yen-Chia Hsu, and C. Lee Giles. 2020. CODA-19: Using a non-expert crowd to annotate research aspects on 10,000+ abstracts in the COVID-19 open research dataset. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*, Online. Association for Computational Linguistics.

Ryan Kiros, Yukun Zhu, Russ R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought vectors. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Yuta Kobayashi, Masashi Shimbo, and Yuji Matsumoto. 2018. Citation recommendation using distributed representation of discourse facets in scientific articles. In *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries*, JCDL '18, page 243–251, New York, NY, USA. Association for Computing Machinery.

Shuo Liang, Wei Wei, Xian-Ling Mao, Fei Wang, and Zhiyong He. 2022. BiSyn-GAT+: Bi-syntax aware graph attention network for aspect-based sentiment analysis. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1835–1848, Dublin, Ireland. Association for Computational Linguistics.

Lajanugen Logeswaran and Honglak Lee. 2018. An efficient framework for learning sentence representations. In *International Conference on Learning Representations*.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605.

Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. 2014. A SICK cure for the evaluation of compositional distributional semantic models. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 216–223, Reykjavik, Iceland. European Language Resources Association (ELRA).

Malte Ostendorff, Till Blume, Terry Ruas, Bela Gipp, and Georg Rehm. 2022. Specialized document embeddings for aspect-based similarity of research papers. In *Proceedings of the 22nd ACM/IEEE Joint Conference on Digital Libraries*, JCDL '22, New York, NY, USA. Association for Computing Machinery.

Malte Ostendorff, Terry Ruas, Till Blume, Bela Gipp, and Georg Rehm. 2020a. Aspect-based document similarity for research papers. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6194–6206, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Malte Ostendorff, Terry Ruas, Moritz Schubotz, Georg Rehm, and Bela Gipp. 2020b. Pairwise multi-class document classification for semantic relations between wikipedia articles. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020*, JCDL '20, page 127–136, New York, NY, USA. Association for Computing Machinery.

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.

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

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525, Online. Association for Computational Linguistics.

Phillip Schneider, Tim Schopf, Juraj Vladika, Mikhail Galkin, Elena Simperl, and Florian Matthes. 2022. A decade of knowledge graphs in natural language processing: A survey. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 601–614, Online only. Association for Computational Linguistics.

Tim Schopf, Karim Arabi, and Florian Matthes. 2023a. Exploring the landscape of natural language processing research.

Tim Schopf, Daniel Braun, and Florian Matthes. 2021. Lbl2vec: An embedding-based approach for unsupervised document retrieval on predefined topics. In *Proceedings of the 17th International Conference on Web Information Systems and Technologies - WEBIST*, pages 124–132. INSTICC, SciTePress.

Tim Schopf, Daniel Braun, and Florian Matthes. 2023b. Evaluating unsupervised text classification: Zero-shot and similarity-based approaches. In *Proceedings of the 2022 6th International Conference on Natural Language Processing and Information Retrieval*, NLPIR '22, page 6–15, New York, NY, USA. Association for Computing Machinery.

Tim Schopf, Daniel Braun, and Florian Matthes. 2023c. Semantic label representations with lbl2vec: A similarity-based approach for unsupervised text classification. In *Web Information Systems and Technologies*, pages 59–73, Cham. Springer International Publishing.

Tim Schopf, Simon Klimek, and Florian Matthes. 2022. Patternrank: Leveraging pretrained language models and part of speech for unsupervised keyphrase extraction. In *Proceedings of the 14th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2022) - KDIR*, pages 243–248. INSTICC, SciTePress.

Tim Schopf, Dennis Schneider, and Florian Matthes. 2023d. Efficient domain adaptation of sentence embeddings using adapters.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: A free collaborative knowledgebase. *Commun. ACM*, 57(10):78–85.

Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Doug Burdick, Darrin Eide, Kathryn Funk, Yannis Katsis, Rodney Michael Kinney, Yunyao Li, Ziyang Liu, William Merrill, Paul Mooney, Dewey A. Murdick, Devvret Rishi, Jerry Sheehan, Zhihong Shen, Brandon Stilson, Alex D. Wade, Kuansan Wang, Nancy Xin Ru Wang, Christopher Wilhelm, Boya Xie, Douglas M. Raymond, Daniel S. Weld, Oren Etzioni, and Sebastian Kohlmeier. 2020. CORD-19: The COVID-19 open research dataset. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*, Online. Association for Computational Linguistics.

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, Melbourne, Australia. Association for Computational Linguistics.

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, Online. Association for Computational Linguistics.

Wenxuan Zhang, Xin Li, Yang Deng, Lidong Bing, and Wai Lam. 2021. Towards generative 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 2: Short Papers)*, pages 504–510, Online. Association for Computational Linguistics.