A Hybrid Approach To Aspect Based Sentiment Analysis Using Transfer Learning

Gaurav Negi[1], Rajdeep Sarkar[2]*, Omnia Zayed[1], Paul Buitelaar[1]
[1]Insight SFI Centre for Data Analytics, University Of Galway
[2]Yahoo Research
{gaurav.negi, omnia.zayed, paul.buitelaar}@insight-centre.org
rajdeep.sarkar@yahooinc.com

Abstract

Aspect-Based Sentiment Analysis (ABSA) aims to identify terms or multiword expressions (MWEs) on which sentiments are expressed and the sentiment polarities associated with them. The development of supervised models has been at the forefront of research in this area. However, training these models requires the availability of manually annotated datasets which is both expensive and time-consuming. Furthermore, the available annotated datasets are tailored to a specific domain, language, and text type. In this work, we address this notable challenge in current state-of-the-art ABSA research. We propose a hybrid approach for Aspect Based Sentiment Analysis using transfer learning. The approach focuses on generating weakly-supervised annotations by exploiting the strengths of both large language models (LLM) and traditional syntactic dependencies. We utilise syntactic dependency structures of sentences to complement the annotations generated by LLMs, as they may overlook domain-specific aspect terms. Extensive experimentation on multiple datasets is performed to demonstrate the efficacy of our hybrid method for the tasks of aspect term extraction and aspect sentiment classification.

Keywords: Aspect Based Sentiment Analysis, Syntactic Parsing, large language model (LLM)

1. Introduction

Aspect-Based Sentiment Analysis (ABSA) (Liu, 2012) refers to the task of identifying the aspects of the entities and their associated sentiments from a given text sequence. ABSA comprises two fundamental tasks: Aspect Term Extraction (ATE) and Aspect Sentiment Classification (ASC). The significance of fine-grained ABSA becomes apparent when different sentiments are articulated concerning distinct attributes of an entity, as highlighted by Hu and Liu (2004). An Aspect Term is a single- or multi-word expression within the text that serves to describe a specific aspect or attribute of the entity upon which sentiment is being expressed. When applied to a collection of review sentences, ATE involves the identification of all aspect terms or opinion targets contained within each sentence. Subsequently, ASC is concerned with the classification of sentiments associated with each of the aspect terms that were identified during the ATE process.

For instance, consider the sentence: "*I liked the service and the staff, but not the food.*" This sentence conveys nuanced sentiments pertaining to specific aspects. In particular, the aspect terms service and staff are associated with a positive sentiment, while the aspect term food carries a negative sentiment within the given context. It is crucial to emphasise the domain specificity of these aspect terms and the intricate relationships that exist between them as it enables a better understanding about the product, instead of directly analysing the sentiment of the text as a whole. In the example provided, the terms service, staff, and food collectively suggest that the text originates from a restaurant review. A model trained for a different domain, such as "Electronics" would struggle to identify these domain-specific terms in the "restaurant" domain, thereby leading to sub-optimal performance in sentiment analysis tasks. This underscores the requirement for domain-specific ABSA frameworks to enhance performance within specific domains. Nevertheless, constructing domain-specific datasets entails significant costs in terms of iterative efforts and the involvement of specialised personnel. Consequently, there arises a demand for the development of domain-specific ABSA systems that can operate effectively without the necessity of manually annotated training datasets.

We initiate our investigation by employing syntactic dependencies to identify aspect terms within a domain-specific context. However, our findings reveal that Machine Learning models trained on generic datasets are not universally applicable across various domains. Consequently, there arises a necessity for the domain adaptation of machine learning methodologies to address this task. Machine learning solutions generally treat ATE and ASC as supervised tasks. As supervised approaches rely on annotated datasets, their utility for ABSA in diverse domains is inherently limited. In response to this limitation, unsupervised approaches have been explored to a limited extent

*Work done as a PhD student at University of Galway.

LREC-COLING 2024, pages 647–658
20-25 May, 2024.

in ABSA, some of which utilise pseudo-labeling as an interim technique within their framework (Giannakopoulos et al., 2017; Wu et al., 2018a).

Large Language Models (LLMs) (Brown et al., 2020) have ushered in a transformative era in the realm of Natural Language Processing (NLP). These models undergo pre-training on extensive unlabeled corpora and subsequent fine-tuning across various labeled downstream tasks. Research (Wei et al., 2022; Chung et al., 2022) has demonstrated that fine-tuning over a multitude of tasks, each with specific instructions, enhances the capacity of a model to generalise to novel tasks that were not encountered during pre-training.

In our research, we leverage LLMs to address the ATE task within a domain-specific context. Notably, we discover that while LLMs exhibit considerable power, they tend to exhibit limitations in capturing aspect terms with a higher recall rate, leading to suboptimal performance. This observation prompts a fundamental question: "Can LLMs fine-tuned on a source domain, where abundant annotated training data is available, be effectively adapted to another target domain that lacks such annotated resources?". To investigate this, we employ a transfer-learning approach, fine-tuning an LLM in the source domain (Gadgets and Social Media) and subsequently evaluating its performance on the target domain (Restaurant and Laptop). Our findings reveal that while the model adapts to address the in-domain ATE task, it still grapples with domain-specific challenges inherent to the target domains.

In response to these challenges, we introduce a hybrid solution designed to generate synthetic, domain-specific Aspect-Based Sentiment Analysis (ABSA) datasets through transfer learning. We leverages the LLM that has been fine-tuned for the ATE task, utilising it to generate domain-specific training data. This dataset is then enriched through the incorporation of syntactic dependencies, which enhances the recall of identified aspect terms. Subsequently, this refined training dataset is employed to conduct the fine-tuning of another LLM, resulting in the development of a domain-grounded ABSA model. In summary, our main contributions are as follows:

- Method for domain adaptation with Transfer Learning for ATE, using LLMs. Including a hybrid method of annotating training datasets for ATE.
- Evaluating the performance of GPT-3.5-Turbo and Flan-T5 in zero-shot setting for the task of ATE and ASC.
- Flan-T5 model fine-tuning and investigation of performance with qualitative and quantitative comparative analysis.

2. Related Work

The conception of ABSA entailed the feature-based summarisation of customer reviews of products sold online (Hu and Liu, 2004). Features of the product on which the customers have expressed their opinions. Subsequently, a distinction was made between feature keywords and opinion words (Zhuang et al., 2006). A holistic algorithm for feature-based sentiment analysis utilising context was introduced by (Ding et al., 2008). Over time, this evolved into the research area of ABSA, where the tasks of ATE and ASC have been explored with different methods.

Rule Based Methods for ABSA: These approaches utilised POS tagging, opinion lexicons (Liu et al., 2005) and the syntactic dependency parser to identify the syntactic relationship between opinion words and aspect terms. Almost all initial ABSA methods followed a rule-based approach (Hu and Liu, 2004; Zhuang et al., 2006). Double Propagation (Qiu et al., 2011, 2009) introduced methods for the dynamic expansion of opinion words. Automatic methods for effective dependency parsing rule selection were also introduced (Liu et al., 2015).

Supervised Machine Learning Methods: ATE can be formulated as a sequence labelling task and ASC as a classification task. Initial methods of note performed feature enrichment using a large unlabelled corpus (Amazon product reviews, Yelp Reviews) before using ML models like CRF (Wang et al., 2016; Toh and Wang, 2014) and SVM (Clercq et al., 2015). Neural Network architectures that encode and utilise the sequential information with Long Short-Term Memory (LSTM) neural networks, Convolutional Neural Networks (CNN) and Attention were extensively used for ABSA. Some models also successfully utilised the relatedness of ATE and aspect-based sentiment analysis by multi-task learning with hard parameter sharing (Ruder, 2017) in the deep learning model (Xue et al., 2017a; Li et al., 2018). The current state-of-the-art results are given by (Scaria et al., 2023), which utilises the LLM introduced by (Wang et al., 2022). It outperforms the previous state-of-the-art which was BERT-based models (Yang and Li, 2022), (Zhang et al., 2022).

Unsupervised Machine Learning Methods: The objective of unsupervised ABSA methods is to avoid the usage of the annotated datasets. These approaches often are a combination of rule-based ABSA with heuristics to annotate datasets(Giannakopoulos et al., 2017; Wu et al., 2018b; Karaoglan and Findik, 2022).

Supervised machine learning approaches require domain-specific annotated datasets. Data annotation is expensive and time-consuming. Un-

Source	Domain	Training	Testing
SemEval (2014)	Laptop Restaurant	3041 3045	800 800
Twitter Sentiments	Open Domain	6248	692
Gadget Reviews	Electronic Products	2099	-

Table 1: ABSA datasets

Dependency Relation	Aspect Term(s)	Opinion Word
AT—DEP—O	AT	O
AT—DEP_1—M—DEP_2—O	AT	O
AT_1—DEP_3—AT_2	AT_1, AT_2	—

Table 2: Syntactic Dependencies for ATE

supervised methods are often complex and restrictive in their performance. Rule-based approaches are not as effective as the approaches that utilise contextual information effectively, i.e., LSTM, Transformers, BERT, T5. To the best of our knowledge domain specificity and similarity of the sources of existing annotated datasets (reviews) exists in ABSA. In this paper, we describe an ABSA method based on transfer learning and domain adaptation. The approach focuses on generating weakly-supervised annotations by exploiting the strengths of both LLMs and traditional syntactic dependencies. We use LLMs because of their versatility, understanding of natural instruction and the generalised nature of learning through instruction fine-tuning (Brown et al., 2020; Wei et al., 2022).

3. Datasets

We utilise the Restaurant and Laptop reviews dataset (Pontiki et al., 2014), Twitter Sentiments dataset (Dong et al., 2014) and the Garget reviews dataset (Liu et al., 2015) to conduct our experiments on different tasks. The Restaurant and Laptop reviews dataset consists of reviews categorised into two categories: *restaurant* and *laptop*. The Gadget reviews dataset contains product reviews spanning three categories: *speakers*, *computers* and *router*. Each review in the two datasets is annotated with aspect terms and their corresponding sentiment polarity. In the restaurant category, the sentences are also annotated with aspect category information and aspect category polarity. The Twitter Sentiments dataset provides manually annotated tweets for target-dependent sentiment analysis. Each tweet is annotated with a target term and its associated sentiment. The size of the training and test split of the different datasets is detailed in Table 1.

4. Methodology

In this section, we begin with a formal definition of the problem statement. Subsequently, we delve into the process of creating domain-specific ATE dataset utilising syntactic dependencies. We proceed to outline the experimental setup employed for assessing LLMs in the context of the ATE task in a zero-shot learning scenario. Following this, we describe the configuration for domain transfer learning as applied to the Aspect-Based Sentiment Analysis (ABSA) task within LLMs. Finally, we describe our proposed methodology, designed to facilitate the training of domain-specific ABSA models without necessitating the presence of manually annotated training data.

4.1. Problem Formulation

Given the input text $x = x_1, x_2, ..., x_t$, where x_i represents the i^{th} word in the input, ATE pertains to the identification of single- or multi-word terms within x that convey sentiments. Let $a_j = x_k, x_{k+1}, ..., x_m$ denote an aspect term, and ASC involves the determination of the sentiment polarity s_j associated with a_j. The sentiment s can assume one of the values from the set positive, negative, neutral, denoted as $s \in \{positive, negative, neutral\}$.

4.2. Syntactic Dependencies for ATE

Exploiting the syntactic dependency structures have been an integral part of the many ABSA approaches (Zhuang et al., 2006; Qiu et al., 2011; Giannakopoulos et al., 2017). The core concept underlying these approaches is the recognition of multiple syntactic relations that establish connections between opinion words and aspect terms.

We extract Noun Phrases (NP) from the text, considering them as potential aspect term candidates. To refine this candidate set, we perform a pruning step to eliminate all stop words and opinion words. Subsequently, we select the candidates that follow the syntactic dependency structures (Dozat et al., 2017) described in Table 2 and designate them as aspect terms. AT, AT_1, and AT_2 are NPs, which were obtained by chunking. O are the opinion words with the specific POS tags. The potential opinion words mentioned in the table O are the words that are Adjectives(JJ), verbs (VB) and adverbs (RB). DEP, DEP_1, DEP_2 and DEP_3 are the placeholders for universal dependency (UD) (Nivre et al., 2016) tags. They are used for defining syntactic relationships within the sentence. Their values are specified as follows:

$$DEP \in (amod, nsubj, xcomp, obl, obj, nmod, dep)$$
$$DEP_1 \in (amod, nsubj, nmod)$$
$$DEP_2 \in (amod, nsubj, xcomp, advmod, nmod)$$
$$DEP_3 \in (conj)$$

M is a placeholder for the words which are indirectly related to Aspect Term and Opinion which takes on the Noun (NN), Verb (VB) and Adverb (RB) POS tag values. For e.g. in the following sentence "*I liked the service and the staff, but not the food.*" $AT-DEP-O$ type relationship exists between *service* and *liked*, $AT_1-DEP_3-AT_2$ type relationship between *service* and *staff* , enabling us to extract those two aspect terms.

The syntactic dependency annotation method plays a crucial role in ABSA. However, it is essential to acknowledge its limitations. Due to its lack of specificity, this method often exhibits lower precision than recall. Consequently, a notable issue arises: a significant number of terms are incorrectly identified as aspect terms. Addressing this challenge is vital for improving the accuracy of ATE.

4.3. Zero Shot ATE and ASC

In this section, we describe the zero-shot setup of LLMs for the ATE and the ASC task. In a zero-shot setting, we do not have access to labelled data, hence we leverage prompting (Liu et al., 2021) to obtain aspect terms and their associated sentiments. Using the designated prompt templates, Prompt 1 and Prompt 2, we input text instances denoted as x_i into the LLM alongside the respective prompts. The template Prompt 1 serves as a directive for the LLM, instructing it to discern the pertinent aspect terms within the given context. In tandem, Prompt 2 functions to attribute a specific sentiment polarity to the AT identified in the preceding step. It is crucial to emphasise that this configuration operates in the absence of labeled data, and the LLM remains non-fine-tuned for either the ATE or ASC task.

4.4. Transfer Learning for ATE and ASC

In this configuration, we fine-tune a pre-trained Large Language Model (LLM) using the Twitter Sentiment and Gadget Reviews datasets, specifically for the ATE and ASC tasks. This fine-tuning phase is instrumental in facilitating the adaptation of the LLM to these tasks. Notably, the multi-domain characteristic of the Twitter Sentiment and Gadget Reviews datasets equips the LLM with the capability to generalise its performance across the ATE and ASC tasks. Subsequently, this fine-tuned model is harnessed within a transfer-learning framework for evaluation on domain-specific datasets.

In our approach, we engage in the fine-tuning of the LLM within a text-to-text format framework, wherein both the input and output are represented in textual form. Specifically, for both the ATE and ASC tasks, the LLM operates on input data in the form of prompt templates, yielding outputs in the form of either aspect terms or their corresponding sentiments. This fine-tuning process on multi-domain datasets serves the purpose of task adaptation, leveraging these prompt templates to enhance performance of the LLM.

Input (x_i) = Extract aspect terms from the following input.
input: just watched ps i love you on star movies . i love hilary swank's smile ! </s>
Output (y_i) = hilary swank </s>

Prompt 1: LLM Prompt for ATE

Aspect Term Extraction (ATE): The model takes input x_i and label y_i at the training time. An example scenario is illustrated in Prompt 1.
Aspect Sentiment Classification (ASC): When predicting the sentiment class of an AT in x_i, we utilise the prmpt structure used in Prompt 2. The output label y_i will be the polarity and can take values between (positive, negative or neutral). An example scenario is illustrated in Prompt 2.

Input (x_i) = Given the aspect term and the sentence. Predict if the aspect term in the sentence has a positive, negative or neutral sentiment expressed on it.
aspect term: hilary swank
sentence: just watched ps i love you on star movies. i love hilary swank's smile!</s>
Output (y_i) = positive</s>

Prompt 2: LLM Prompt for ASC

4.5. Weakly Annotating Datasets for ATE

Weakly annotated training datasets are used for adapting model to the training domain.In this section we discuss three methods of automatic annotations of the training text.

4.5.1. Annotation With LLMs

In our approach, we employ fine-tuned LLMs as described in Section 4.4 to generate domain-specific data annotations. Specifically, we utilise training data in templated forms, guided by prompts referenced as 1 and 2, and generate annotations for each input text. This annotated training dataset based on LLMs serves as the foundation for training domain-adapted models. Notably, our hybrid annotation method also leverages this dataset.

The entire process is graphically depicted in Figure 1. Within this process, an out-of-domain dataset is utilised to fine-tune a pre-trained LLM and this fine-tuning is carried out separately for the ATE and ASC task. The resulting out-of-domain

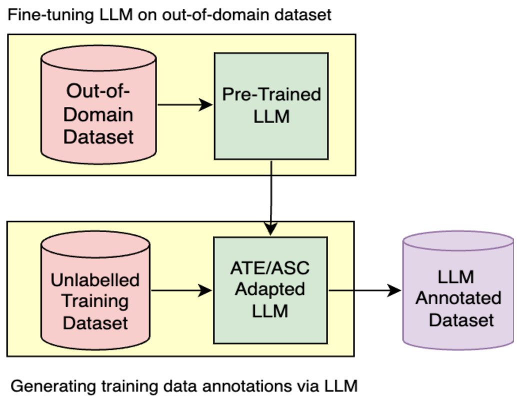

Figure 1: Automatic annotations with LLM

adapted models are subsequently employed to annotate the in-domain training dataset, which, in turn, is used to fine-tune domain-specific ATE and ASC models.

4.5.2. Hybrid Annotation Method

Aspect terms are known to exhibit a high degree of domain-dependence. Consequently, annotations generated by out-of-domain fine-tuned LLMs fail to fully grasp certain highly domain-specific aspect terms. To address this limitation, we strategically enhance LLM annotations by strategically incorporating the syntactic dependency annotation method.

We introduce a hybrid annotation approach that aims to systematically integrate the strengths of both methods. The syntactic dependency annotation method, known for its high recall, complements the LLM annotations, which are distinguished by their high precision. By carefully merging these two approaches, we aim to strike a balance that allows us to retain the precision advantages of the LLM-annotated training dataset while also benefiting from the broader coverage and recall offered by the syntactic dependency annotation method.

Figure 2 shows an overview of our hybrid annotation method for ATE. We begin with a dataset of unlabelled training text (N) with n sentences. The first step is to obtain high-precision annotations from a fine-tuned LLM (Flan-T5-Base-ATE) that is trained on a domain which is different from the domain of N. Once the dataset N has been labelled by the LLM, we consider a subset M of N, that contains at least one annotation by the LLM. From the remaining sentences in N that are not present in M, we select those sentences that have a high semantic similarity to sentences in M. The dataset of sentences thus selected be represented by S. Then we extend the annotations for (M + S) with the Syntactic dependency method.

Candidate Selection for annotation extension: We start with n data points of LLM annotated dataset, where N is their vector of texts from the dataset. We observe that the LLM annotation (Figure 1) has a high precision and leaves out about 50% of the training corpus without annotations. This situation encourages us to explore the integration of annotation methods. For this purpose, we consider two distinct categories of candidates from N for label extension :

1. The instances of N that have at least one aspect term generated. This split of m texts selected from the initial n data points is represented as the vector M.

2. The instances which have no annotations but are semantically similar to the annotated text (s) also from n data points. This similar text split of size s the data set with no annotation is represented as text vector S.

We form these splits on the basis semantic similarity using sentence encoder (Reimers and Gurevych, 2019) to get d dimensional vector representations of the sentences. We encode M to vector $V_M \in \mathbb{R}^{m\times d}$. We calculate the mean vector $\mu_M \in \mathbb{R}^{1\times d}$ from V_M to select s most similar text from Q where $Q = N - M$:

$$V_M = \text{SentenceEncoder}(M) \tag{1}$$

$$\mu_M = \frac{\sum_{j=0}^{m} M_j}{m} \tag{2}$$

We select the s most similar texts from Q using the cosine similarity relative to μ_M. To do so, we use the same sentence encoder we use with M to get V_M and encode Q to V_Q.

$$V_Q = \text{SentenceEncoder}(Q) \tag{3}$$

We then calculate the cosine similarity of all individual encoded text in V_Q, resulting in a similarity vector Y which has 1:1 mapping with V_Q.

$$\begin{gathered} Y = \text{CosineSim}(V_Q, \mu_M) \\ \text{where, CosineSim} : V_Q \mapsto Y \end{gathered} \tag{4}$$

Cut-off fraction (cf) is the hyper-parameter that helps in calculating the cut-off value (C_x) for the similarity for a value from Q to be included in S. This cut-off value is calculated in terms of mean (μ_y) and standard deviation (σ_y) values of similarities Y.

$$C_x = \mu_y + \sigma_y * cf \tag{5}$$

We then use C_x as a threshold for similarity and select S text vector from Q, which along with M is passed through Syntactic Dependency annotation to extend the labels as shown in Figure 2.

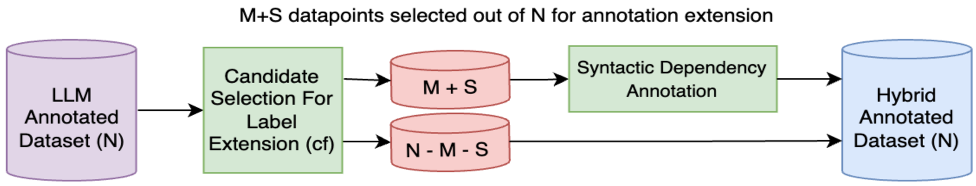

Figure 2: Hybrid annotations

$$X = \begin{cases} S & , \text{CosineSim}(V_Q, \mu_M) > C_x \\ R & , \text{CosineSim}(V_Q, \mu_M) < C_x \end{cases} \quad (6)$$

After the cutoff filtering, we have S with no aspect terms and M texts with at least one aspect term. We combine $S + M$ and extract aspect terms with syntactic dependency structure as discussed in Section 4.2. If there are overlapping token in extraction aspect terms for M, we ignore the Dependency generated aspect terms in favour of LLM generated ones. At the end of this process we combine the newly annotated (S, M) with R non-annotated split to get Hybrid Annotated Dataset.

5. Experiments

We start this section with the implementation details of the experiments. In Section 6, we perform a comparative analysis of the results.

5.1. Implementation Details

For our experimental setup, we utilise for the Flan-T5-Base[1] language model as the LLM for our tasks. Adhering to the methodology outlined by Chung et al. (2022), we employ Adafactor (Shazeer and Stern, 2018) as our chosen optimiser and conduct training for 32 epochs using an NVIDIA Tesla V100 GPU. The best performing model out of 32 epochs is selected. The learning rate is set at 0.0003. Our training configuration involves a batch size of 4 with a gradient accumulation step size of 4, effectively simulating a batch size of 16. The sentence encoder utilised in the hybrid annotation method, as detailed in Section 4.5.2, is the all-MiniLM-L6-v2[2]. This encoder was selected from the sentence encoder leader-board. All hyper-parameters were determined through grid search.

6. Results And Discussion

We compare our ATE/ASC approaches with zero-shot LLM predictions. To establish the upper bounds of performance metrics, we compare with the performance of a supervised model. Unlike our weakly supervised method, which uses automatically annotated training data, the supervised model relies on gold labels. The results of the ATE experiments are presented in Table 4 and 5. We evaluate ATE models in two steps:

1. The first set of evaluations is testing our models and labelling methods on the test set directly. Here, we test the efficiency of our data generation method on test data without utilising training data for domain-adaptation. These results are presented in Table 3. **Syntactic Dependency** method and **Flan-T5-Base (Zero Shot)** are the approaches which does not require any training while **Flan-T5-Base-ATE** uses Flan-T5-Base which has been fine-tuned on Gadget Reviews and Twitter sentiments.

2. In the final step we evaluate the models that were fine-tuned on different annotations for the training dataset. These results are in Table 4.**Flan-ATE-GOLD** is fine-tuned on gold ATE labels. Training data annotated with Flan-T5-ATE-Base generation, using syntactic dependencies and hybrid method is used to fine-tune Flan-T5-Base to obtain **Flan-ATE-DOM-ADAPT**, **Flan-ATE-DEP** and **Flan-ATE-HYBRID** respectively.

The results pertaining to Aspect-Based Sentiment Classification (ASC) are presented in Table 5. All models utilise Prompt 2 for predictions and fine-tuning. The evaluation metrics are macro-averaged across sentiment classes, with any aspects exhibiting conflicting predictions removed from the test dataset. Each prediction entails using the text and aspect terms from the test data to predict sentiments. We perform sentiment predictions without fine-tuning on both GPT-3.5-Turbo and Flan-T5-Base for zero-shot inference. Flan-T5-TL is Flan-T5-Base fine-tuned on Twitter Sentiments and gadget data. Flan-T5-DOM-ADAPT is trained on in-domain generated polarities using Flan-T5-TL, while Flan-T5-GOLD is trained on the original gold sentiment labels of the training data.

6.1. Quantitative Evaluation

ATE Results: For the zero-shot evaluation as seen in Table 3, Flan-T5-Base (Zero Shot) has considerably low accuracy scores, meaning that it does

[1] https://huggingface.co/google/flan-t5-base
[2] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

Method	Fine-Tuned	Laptop Review 2014			Restaurant review 2014		
		Precision	Recall	F1	Precision	Recall	F1
Syntactic Dependency	No	29.54	37.23	32.94	48.53	52.65	50.50
Flan-T5-Base (Zero Shot)	No	12.71	17.07	14.57	19.83	15.09	17.14
Flan-T5-Base-ATE	Yes	53.60	51.53	52.54	67.93	39.45	49.91

Table 3: ATE annotation efficiency on test splits

Model	Type of Annotations	Laptop Review 2014			Restaurant review 2014		
		Precision	Recall	F1	Precision	Recall	F1
GPT-3.5-TURBO	–	22.62	70.61	34.26	36.07	76.46	49.02
Flan-ATE-DOM-ADAPT	Flan-T5-ATE-Base	66.07	45.84	54.13	77.95	43.68	55.99
Flan-ATE-DEP	Syntactic Dependency-based	28.94	47.69	36.02	54.83	61.51	57.98
Flan-ATE-HYBRID	Hybrid	47.85	60.15	53.50	55.57	69.54	61.77
FLan-ATE-GOLD	Gold Labels	86.22	81.84	83.97	87.04	84.81	85.91

Table 4: ATE efficiency after utilising training datasets with different annotation strategies

a poor job of annotating the test data for aspect terms. The score syntactic dependency method of annotation is higher than zero-shot. We see that the model (Flan-T5-ATE-Base), fine-tuned on the Gadget Reviews and Twitter Sentiments, shows an improvement in the annotation generation capabilities.

The results presented in Table 4 pertain to models that underwent fine-tuning using training data annotated through various annotation methods, signifying the domain adaptation process for ATE. The upper performance bound is established by a model that was fine-tuned using gold ATE labels in a purely supervised fashion. Notably, the specifics of the data utilised for training GPT-3.5-Turbo are not disclosed, which is evident in the observed results. GPT-3.5-Turbo exhibits notably high recall but lower precision, resulting in an F1-score that reflects its tendency to extract a multitude of aspect term candidates with relative inaccuracy. Models trained with syntactic dependency annotation achieve comparatively favorable performance in only one of the domains, albeit still exhibiting suboptimal precision values. Conversely, when we train a model with annotations generated by Flan-T5-ATE-Base, a model adapted for the task using out-of-domain data, we observe a more balanced performance across both domains.

Incorporating the hybrid annotation method, our primary objective is to amalgamate annotation techniques characterised by high precision (Flan-T5-ATE-Base) and high recall (syntactic dependency annotations) in order to enhance the overall F1 score of the fine-tuned model (Flan-ATE-HYBRID). It is evident that, in the restaurant domain, this approach surpasses the F1-score in comparison to the previously discussed methods. Conversely, in the case of Laptop Reviews, when the model is trained using hybrid annotation for the training dataset, we observe superior performance in terms of F1-score compared to the model exclusively trained on syntactic dependency annotations and GPT-3.5-TURBO. Nevertheless, the F1-score remains slightly lower than that achieved by the ATE model trained on annotations generated by Flan-T5-ATE-Base.

ASC results: The results pertaining to ASC in Table 5 reveal a notable zero-shot performance of the Flan-T5-BASE model. We posit that this zero-shot performance can be attributed to the efficiency of Flan fine-tuning tasks, as demonstrated by Chung et al. (Chung et al., 2022). Flan-T5-TL, fine-tuned on out-of-domain data, exhibits improvements over the zero-shot performance. GPT-3.5-Turbo, a model of considerable scale and fine-tuning tasks, outperforms Flan-T5-TL, possibly due to its larger size and diverse fine-tuning objectives. Moreover, we observe that the domain-adapted model (Flan-T5-DOM-ADAPT), which underwent fine-tuning on in-domain auto-annotated training data using Flan-T5-TL, outperforms GPT-3.5-TURBO. Furthermore, the model Flan-T5-GOLD, fine-tuned on gold annotated training data, demonstrates robust evaluation metrics. Making it a good upper bound reference for the performance of our ASC models.

6.2. Qualitative Evaluation

6.2.1. ASC Results

In Figure 3, we present a comparative analysis of class-wise correct predictions made by models using the gold labels from the test set across both domains. Flan-T5-TL and Flan-T5-DOM-ADAPT represent our fine-tuned models, the details of which have been previously elucidated. Notably, the models exhibit a certain level of difficulty in accurately predicting the *neutral* labels. GPT-3.5-Turbo and the domain-adapted Flan-T5-DOM-ADAPT excel in correctly predicting the *neutral* class. This difficulty in predicting the *neutral* class has an impact on the macro-averaged metrics. However, when predicting *positive* and *negative* classes, Flan-T5-TL, Flan-T5-DOM-ADAPT, and Flan-T5-Base (zero-shot) showcase comparable, if not superior, perfor-

Model	Type of Annotations	Laptop Review 2014 Precision	Laptop Review 2014 Recall	Laptop Review 2014 F1	Restaurant review 2014 Precision	Restaurant review 2014 Recall	Restaurant review 2014 F1
GPT-3.5-TURBO	–	74.42	75.69	74.95	71.53	74.47	72.87
Flan-T5-BASE	None (Zero Shot)	63.89	67.87	63.67	63.09	66.97	63.54
Flan-T5-TL	(Twitter Sentiments + gadget review)	74.65	70.04	68.74	70.76	70.15	70.16
Flan-T5-DOM-ADAPT	(Flan-T5-Base-InDomain-Gen)	80.76	79.81	79.55	74.35	75.19	74.72
FLan-T5-GOLD	Gold Labels	78.71	80.47	78.91	84.73	80.20	82.23

Table 5: ASC results

Text	GPT-3.5-TURBO	Flan-ATE-DOM-ADAPT	Flan-ATE-HYBRID	GOLD
Two wasted steaks – what a crime!	steaks, wasted, crime	-	steaks	steaks
You will be very happy with the experience.	experience, happiness	experience	experience	-
We didn't know if we should order a drink or leave?	drink, leave	-	drink	drink
The room is a little plain, but it's difficult to make such a small place exciting and I would not suggest that as a reason not to go.	room, plain, small size, excitement	room	room, place	room, place

Table 6: ATE Prediction examples

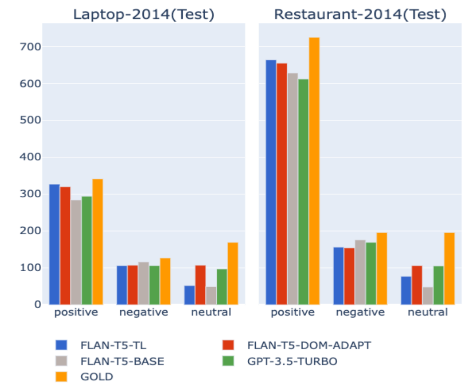

Figure 3: Class-wise Analysis Of ASC models

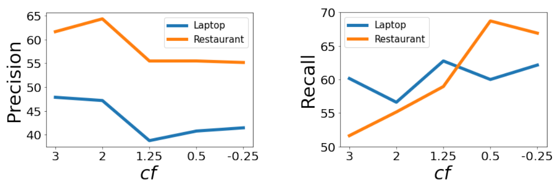

Figure 4: Impact of cf on ATE Hybrid Models

mance compared to GPT-3.5-TURBO.

In Table 6, we present some key results pertaining to ATE. GPT-3.5-TURBO exhibits a high number of false positive aspect terms, which can be attributed to its significantly higher recall but lower precision. Conversely, Flan-ATE-DOM-ADAPT demonstrates high precision, which leads to missed aspect terms. In contrast, Flan-ATE-HYBRID strikes a balance between precision and recall, resulting in occasional false extractions but fewer compared to GPT-3.5-TURBO. This balance contributes to its higher F1 score.

6.2.2. Impact of CF (cutoff fraction) on Performance

In Section 4.5.2, Equation 5 incorporates the cutoff fraction (cf) to determine the similarity cut-off (C_x) for the hybrid annotation method. A higher value of cf corresponds to an elevated similarity cut-off, resulting in a more stringent criterion for data points to be labeled using syntactic dependency as the value of C_x decreases.

In Figure 4, we illustrate the effective control of the precision-recall trade-off achieved through the manipulation of the parameter cf, which governs the annotation splits and contributes to more efficient learning. Lowering the value of cf results in a reduction of the cut-off similarity, thereby allowing a greater number of data points with diminishing similarity to undergo label extension via syntactic dependency structures. Notably, we observe an ascending trend in recall values as cf is decreased. It is worth highlighting an intriguing trend in precision: initially, there is a sharp decline, but with a well-balanced inclusion of segments and exclusion of segments in the annotation process, we witness a subsequent increase in precision. This suggests the potential for further improvements in model performance when operating with larger unannotated datasets.

7. Conclusion

In our work, we undertake a systematic analysis of the performance of non-learnable syntactic dependencies and the application of transfer learning in Large Language Models (LLMs) for the Aspect-Term Extraction (ATE) task in Aspect-Based Sentiment Analysis (ABSA). Subsequently, we introduce an innovative hybrid framework for generating domain-specific datasets by leveraging task-adapted LLMs and syntactic dependencies for both ATE and Aspect-Based Sentiment Classification (ASC) tasks in ABSA. We execute comprehensive experiments across multiple datasets to illustrate the robust empirical performance of our proposed approach compared to other baseline methodologies. In terms of future work, we envisage extending this research to encompass multilingual settings and investigating the utilization of encoder-only architectures.

8. Limitations

We utilise syntactic dependencies and LLMs for generating synthetic dataset. However, LLMs and syntactic dependencies have limited capabilities in a low-resource setting. In the future, we intend to address this limitation. We would also like to extend the scope of our experiments to more open-sourced LLMs and explore the utilisation of encoder-only architecture too.

9. Acknowledgements

This work was conducted with the financial support of the Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289_P2 (Insight_2) and was also supported by funding from the Irish Research Council (IRC) for the Postdoctoral Fellowship award GOIPD/2023/1556.

10. Bibliographical References

Galen Andrew and Jianfeng Gao. 2007. Scalable training of L1-regularized log-linear models. In *Proceedings of the 24th International Conference on Machine Learning*, pages 33–40.

Kamil Bennani-Smires, Claudiu Musat, Andreea Hossmann, Michael Baeriswyl, and Martin Jaggi. 2018. Simple unsupervised keyphrase extraction using sentence embeddings. In *Proceedings of the 22nd Conference on Computational Natural Language Learning, CoNLL 2018, Brussels, Belgium, October 31 - November 1, 2018*, pages 221–229. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Caroline Brun, Diana Nicoleta Popa, and Claude Roux. 2014. XRCE: hybrid classification for aspect-based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation, SemEval@COLING 2014, Dublin, Ireland, August 23-24, 2014*, pages 838–842. The Association for Computer 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.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Y. Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *CoRR*, abs/2210.11416.

Orphée De Clercq, Marjan Van de Kauter, Els Lefever, and Véronique Hoste. 2015. LT3: applying hybrid terminology extraction to aspect-based sentiment analysis. In *Proceedings of the 9th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2015, Denver, Colorado, USA, June 4-5, 2015*, pages 719–724. The Association for Computer Linguistics.

Xiaowen Ding, Bing Liu, and Philip S. Yu. 2008. A holistic lexicon-based approach to opinion mining. In *Proceedings of the International Conference on Web Search and Web Data Mining, WSDM 2008, Palo Alto, California, USA, February 11-12, 2008*, pages 231–240. ACM.

Athanasios Giannakopoulos, Claudiu Musat, Andreea Hossmann, and Michael Baeriswyl. 2017. Unsupervised aspect term extraction with B-LSTM & CRF using automatically labelled datasets. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, WASSA@EMNLP 2017, Copenhagen, Denmark, September 8, 2017*, pages 180–188. Association for Computational Linguistics.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Seattle, Washington, USA, August 22-25, 2004*, pages 168–177. ACM.

Qingnan Jiang, Lei Chen, Ruifeng Xu, Xiang Ao, and Min Yang. 2019. A challenge dataset and effective models 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 2019, Hong Kong, China, November 3-7, 2019*, pages 6279–6284. Association for Computational Linguistics.

Kürsat Mustafa Karaoglan and Oguz Findik. 2022. Extended rule-based opinion target extraction with a novel text pre-processing method and ensemble learning. *Appl. Soft Comput.*, 118:108524.

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 Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, pages 4194–4200. ijcai.org.

Bing Liu. 2012. *Sentiment Analysis and Opinion Mining*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

Jialu Liu, Jingbo Shang, Chi Wang, Xiang Ren, and Jiawei Han. 2015. Mining quality phrases from massive text corpora. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data, Melbourne, Victoria, Australia, May 31 - June 4, 2015*, pages 1729–1744. ACM.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *CoRR*, abs/2107.13586.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2023. MTEB: massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2023, Dubrovnik, Croatia, May 2-6, 2023*, pages 2006–2029. Association for Computational Linguistics.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D. Manning, Ryan T. McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, May 23-28, 2016*. European Language Resources Association (ELRA).

Sinno Jialin Pan and Qiang Yang. 2010. A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.*, 22(10):1345–1359.

Guang Qiu, Bing Liu, Jiajun Bu, and Chun Chen. 2009. Expanding domain sentiment lexicon through double propagation. In *IJCAI 2009, Proceedings of the 21st International Joint Conference on Artificial Intelligence, Pasadena, California, USA, July 11-17, 2009*, pages 1199–1204.

Guang Qiu, Bing Liu, Jiajun Bu, and Chun Chen. 2011. Opinion word expansion and target extraction through double propagation. *Comput. Linguistics*, 37(1):9–27.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

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 2019, Hong Kong, China, November 3-7, 2019*, pages 3980–3990. Association for Computational Linguistics.

Sebastian Ruder. 2017. An overview of multi-task learning in deep neural networks. *CoRR*, abs/1706.05098.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal V. Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Févry, Jason Alan Fries, Ryan Teehan, Teven Le Scao, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M. Rush. 2022. Multitask prompted training enables zero-shot task generalization. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Kevin Scaria, Himanshu Gupta, Saurabh Arjun Sawant, Swaroop Mishra, and Chitta Baral. 2023. Instructabsa: Instruction learning for aspect based sentiment analysis. *CoRR*, abs/2302.08624.

Jingbo Shang, Jialu Liu, Meng Jiang, Xiang Ren, Clare R. Voss, and Jiawei Han. 2017. Automated phrase mining from massive text corpora. *CoRR*, abs/1702.04457.

Noam Shazeer and Mitchell Stern. 2018. Adafactor: Adaptive learning rates with sublinear memory cost. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 4603–4611. PMLR.

Jie Tao and Lina Zhou. 2020a. A weakly supervised wordnet-guided deep learning approach to extracting aspect terms from online reviews. *ACM Trans. Manag. Inf. Syst.*, 11(3):13:1–13:22.

Jie Tao and Lina Zhou. 2020b. A weakly supervised wordnet-guided deep learning approach to extracting aspect terms from online reviews. *ACM Trans. Manag. Inf. Syst.*, 11(3):13:1–13:22.

Tun Thura Thet, Jin-Cheon Na, and Christopher S. G. Khoo. 2010. Aspect-based sentiment analysis of movie reviews on discussion boards. *J. Inf. Sci.*, 36(6):823–848.

Zhiqiang Toh and Wenting Wang. 2014. DLIREC: aspect term extraction and term polarity classification system. In *Proceedings of the 8th International Workshop on Semantic Evaluation, SemEval@COLING 2014, Dublin, Ireland, August 23-24, 2014*, pages 235–240. The Association for Computer Linguistics.

Wenya Wang, Sinno Jialin Pan, Daniel Dahlmeier, and Xiaokui Xiao. 2016. Recursive neural conditional random fields for aspect-based sentiment analysis. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 616–626. The Association for Computational Linguistics.

Wenya Wang, Sinno Jialin Pan, Daniel Dahlmeier, and Xiaokui Xiao. 2017. Coupled multi-layer attentions for co-extraction of aspect and opinion terms. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 3316–3322. AAAI Press.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Anjana Arunkumar, Arjun Ashok, Arut Selvan Dhanasekaran, Atharva Naik, David Stap, Eshaan Pathak, Giannis Karamanolakis, Haizhi Gary Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krima Doshi, Maitreya Patel, Kuntal Kumar Pal, Mehrad Moradshahi, Mihir Parmar, Mirali Purohit, Neeraj Varshney, Phani Rohitha Kaza, Pulkit Verma, Ravsehaj Singh Puri, Rushang Karia, Shailaja Keyur Sampat, Savan Doshi, Siddhartha Mishra, Sujan Reddy A, Sumanta Patro, Tanay Dixit, Xudong Shen, Chitta Baral, Yejin Choi, Hannaneh Hajishirzi, Noah A. Smith, and Daniel Khashabi. 2022. Benchmarking generalization via in-context instructions on 1, 600+ language tasks. *CoRR*, abs/2204.07705.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022. Finetuned language models are zero-shot learners. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Chuhan Wu, Fangzhao Wu, Sixing Wu, Zhigang Yuan, and Yongfeng Huang. 2018a. A hybrid unsupervised method for aspect term and opinion target extraction. *Knowl. Based Syst.*, 148:66–73.

Chuhan Wu, Fangzhao Wu, Sixing Wu, Zhigang Yuan, and Yongfeng Huang. 2018b. A hybrid unsupervised method for aspect term and opinion target extraction. *Knowl. Based Syst.*, 148:66–73.

Wei Xue, Wubai Zhou, Tao Li, and Qing Wang. 2017a. MTNA: A neural multi-task model for aspect category classification and aspect term extraction on restaurant reviews. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing, IJCNLP 2017, Taipei, Taiwan, November 27 - December 1, 2017, Volume 2: Short Papers*, pages 151–156. Asian Federation of Natural Language Processing.

Wei Xue, Wubai Zhou, Tao Li, and Qing Wang. 2017b. MTNA: A neural multi-task model for aspect category classification and aspect term extraction on restaurant reviews. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing, IJCNLP 2017, Taipei, Taiwan, November 27 - December 1, 2017, Volume 2: Short Papers*, pages 151–156. Asian Federation of Natural Language Processing.

Heng Yang and Ke Li. 2022. Pyabsa: Open framework for aspect-based sentiment analysis.

Yichun Yin, Furu Wei, Li Dong, Kaimeng Xu, Ming Zhang, and Ming Zhou. 2016. Unsupervised word and dependency path embeddings for aspect term extraction. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016, New York, NY,* *USA, 9-15 July 2016*, pages 2979–2985. IJCAI/AAAI Press.

Yiming Zhang, Min Zhang, Sai Wu, and Junbo Zhao. 2022. Towards unifying the label space for aspect- and sentence-based sentiment analysis. In *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 20–30. Association for Computational Linguistics.

Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. 2021. A comprehensive survey on transfer learning. *Proc. IEEE*, 109(1):43–76.

Li Zhuang, Feng Jing, and Xiaoyan Zhu. 2006. Movie review mining and summarization. In *Proceedings of the 2006 ACM CIKM International Conference on Information and Knowledge Management, Arlington, Virginia, USA, November 6-11, 2006*, pages 43–50. ACM.

11. Language Resource References

Li Dong, Furu Wei, Chuanqi Tan, Duyu Tang, Ming Zhou, and Ke Xu. 2014. Adaptive recursive neural network for target-dependent twitter sentiment classification. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 2: Short Papers*, pages 49–54. The Association for Computer Linguistics.

Timothy Dozat, Peng Qi, and Christopher D. Manning. 2017. Stanford's graph-based neural dependency parser at the conll 2017 shared task. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, Vancouver, Canada, August 3-4, 2017*, pages 20–30. Association for Computational Linguistics.

Bing Liu, Minqing Hu, and Junsheng Cheng. 2005. Opinion observer: analyzing and comparing opinions on the web. In *Proceedings of the 14th international conference on World Wide Web, WWW 2005, Chiba, Japan, May 10-14, 2005*, pages 342–351. ACM.

Qian Liu, Zhiqiang Gao, Bing Liu, and Yuanlin Zhang. 2015. Automated rule selection for aspect extraction in opinion mining. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015, Buenos Aires, Argentina, July 25-31, 2015*, pages 1291–1297. AAAI Press.

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.