LADy 💃: A Benchmark Toolkit for Latent Aspect Detection Enriched with Backtranslation Augmentation

Farinam Hemmatizadeh
University of Windsor, Canada
hemmatif@uwindsor.ca

Christine Wong
University of Windsor, Canada
wong93@uwindsor.ca

Alice Yu
Vincent Massey Secondary School, Canada
alice.yu@uwindsor.ca

Hossein Fani
University of Windsor, Canada
hfani@uwindsor.ca

ABSTRACT

We present LADy (💃), a python-based benchmark toolkit to facilitate extracting aspects of products or services in reviews toward which customers target their opinions and sentiments. Although there has been a significant increase in aspect-based sentiment analysis, yet the proposed methods' practical implications in real-world settings remain moot for their closed and irreproducible codebases, inability to accommodate datasets from various domains and poor evaluation methodologies. LADy is an open-source benchmark toolkit with a standard pipeline and experimental details to fill the gaps. It incorporates a host of canonical models along with benchmark datasets from varying domains, including unsolicited online reviews. Leveraging an object-oriented design, LADy readily extends to new models and training datasets. The first of its kind, LADy also features review augmentation via natural language backtranslation that can be integrated into the training phase of the models to boost efficiency and improve efficacy during inference. LADy's codebase, along with the installation instructions and case studies on five datasets for seven methods with backtranslation augmentation over ten languages, can be obtained at `https://github.com/fani-lab/LADy`.

1 INTRODUCTION

Customers express opinions and sentiments about products' or services' aspects, also referred to as features, properties, or components, in reviews like *"it's really cheap and you won't regret buying it."* where a customer expresses her opinion *'really cheap'* with *'positive'* sentiments toward *'price'* aspect of a laptop, respectively. Aspects can be explicitly mentioned with a surface form, such as the aspect *'pizza'* in *"pizza was great."*, or they can be implicitly understood (latent) through common social knowledge like *'food'* in *"mine was a little burnt!"*. Aspect extraction finds immediate application in various domains such as sentiment analysis, where it aids in identifying specific aspects or features of products, services, or experiences that customers discuss in their reviews. By extracting these aspects, businesses can assess customer sentiment and make data-driven decisions to improve the quality of their offerings and enhance overall customer satisfaction.

While there has been a surge in aspect extraction research [7, 9, 13, 15, 19, 23, 24], existing methods lack standardized implementations and experimental details, making it challenging to reproduce results consistently across various datasets and domains. For instance, most studies primarily only rely on different versions of `semeval` datasets [11, 24–28, 31], lacking extensibility to incorporate a new dataset and effectively process unsolicited review sourced from social media platforms, such as `Twitter`. This limitation becomes particularly critical when introducing data augmentation methods that require experimental validation across existing and augmented datasets. Moreover, the reported baselines' results are either sourced from other papers or directly extracted from the original papers [3, 22, 27, 32], preventing their reproducibility posing significant challenges for researchers who seek to adopt or improve these techniques. Furthermore, these studies predominantly focus on standard metrics such as precision, recall, and F_1 score. However, this approach may be insufficient if the evaluation strategy requires adaptation to specific scenarios. For example, in situations where the aspect is latent or lacks a clear surface form in reviews, the aspect detection method typically generates a list of the most likely aspects [8]. In such cases, the inclusion of more precise metrics like precision@k or recall@k becomes crucial. This necessitates the capability to integrate new metrics into the evaluation phase, ensuring no need for rewriting the codes for the rest of the pipeline. In response to these challenges, there is a need for a reproducible and user-friendly library. Such a library should be designed to accommodate new datasets and evaluation techniques, facilitating the process for researchers. This would enable more effective comparisons and analysis of results using different leading methods within a unified and accessible toolkit.

LADy is an open-source platform designed to standardize and streamline the analysis of unsolicited online reviews, with a particular focus on the valuable but often overlooked unsolicited customer opinions shared on platforms such as `Google Reviews` and `Twitter`. These unsolicited reviews are characterized by their informality and briefness, often containing textual ellipses and requiring specialized behavior, such as latent aspect detection. To

Table 1: Comparison of existing systems for aspect detection.

features	absa-pytorch[20]	pyabsa[30]	LADy
maintenance	✗	✓	✓
open-source	✓	✓	✓
reproducible	✓	✓	✓
baseline models	static	static	any
datasets	static	format-specific	any
data augmentation	✗	static	any
evaluation metrics	static	static	any
latent aspect detection	✗	✗	✓

this end, LADy offers a standardized implementation environment, benchmark datasets, adaptability to new metrics, an object-oriented structure, and open-source accessibility, making it a valuable toolkit for the natural language processing research community. It promotes reproducibility and fair model comparisons by providing an array of supervised and unsupervised aspect detection methods and empowers researchers to explore different evaluation approaches using different evaluation metrics. Furthermore, LADy improves its aspect detection capabilities through augmentation techniques such as backtranslation augmentation [8] to expand the vocabulary and increase the diversity of aspects detected, further improving the extraction of latent aspects.

In the literature, there are limited open-source toolkits and libraries dedicated to the task of aspect-based sentiment analysis. These have been systematically compared with LADy in terms of various features, including reproducibility, latent aspect detection capability, evaluation strategies, and more, as detailed in Table 1. An outline of the primary flow of LADy for detecting latent aspects enriched with the backtranslation augmentation technique is illustrated in Figure 1.

2 ASPECT DETECTION

Given a collection of reviews $\mathcal{R}$ and a set of aspects $\mathcal{A}$, a review is about a product, service, or feature referred to as an aspect a, belonging to $\mathcal{A}$. For instance, a review r_1 like *"The dessert was the perfect ending!"* explicitly mentions aspect a_1 as *'dessert.'* On the other hand, a review r_2 such as *"mine was a little burnt!"* lacks an explicit surface of the latent aspect a_2 like *'food'* in the review.

Aspects may not necessarily exist with an explicit mention in the review, and the aspect detection methods are mostly originally designed to find the explicit version of aspects. We addressed it by employing a data augmentation technique to come up with new words that can help with finding the latent occurrences of aspects, utilizing the backtranslation method. During the data augmentation phase, the objective is to generate a translated review r' in a specific language l based on an input review r with its associated aspect a. Subsequently, the backtranslated review r'' is created in the source language, which is `English` with its new aspect a''. This new collection of backtranslated reviews $\mathcal{R}''$ is later utilized along with the original collection of reviews $\mathcal{R}$ to empower the existing models to become able to detect the latent aspects as well as the explicit ones.

3 SYSTEM DESIGN AND ARCHITECTURE

We present a comprehensive depiction of the LADy toolkit in Figure 1, which is an advanced system adept at transforming raw reviews from diverse domains into refined abstract review entities with attributes such as aos (aspect, opinion, and sentiment) and authorship details (3.1). An outstanding aspect of LADy is its dynamic adaptability, functioning as a canonical host. It leverages polymorphism to integrate and override functions such as training and inference within the abstract aspect detection class (3.3). Following this, LADy masks explicit aspect labels in reviews to simulate latent aspects, creating datasets to effectively evaluate its latent aspect detection capabilities (3.4). Additionally, LADy applies data augmentation techniques, notably backtranslation across different languages, to strengthen its capacity to recognize both explicit and latent aspects in reviews (3.2). The subsequent sections provide a detailed exploration of each component and layer within the LADy toolkit, highlighting their specific functions and how they collectively enhance the system's overall effectiveness.

3.1 Review Class

Our toolkit is built on a foundational layer known as the data layer, acting as the cornerstone of our architecture. This layer sets up the essential infrastructure for managing reviews by the AbstractReviewClass. The review class diagram, depicted in Figure 2, offers a detailed perspective of this structure. Each review object is structured to include a detailed array of aos triplets, each representing an aspect, opinion, and sentiment for every sentence in the review, as illustrated in Figure 2. Additionally, Figure 2 highlights that a review can contain customized features like timestamps and geographical information, adding significant value for further analytical purposes.

Accompanying the illustrations, Figure 3 presents a code snippet and summary of the review class, detailing its properties and functions. This visualization aids in understanding the class's structure and the roles of its various components. The general aim of the abstract review class is to introduce a unified level of abstraction across diverse review datasets from sources like `twitter`, `semeval`, etc., featuring varied file formats and domains such as restaurants, laptops, and social media. These datasets, whether in `json`, `csv`, or other formats, are streamlined under this model. For instance, `semeval` datasets utilize `xml`-based reviews, whereas `twitter` dataset is formatted as `txt` with different annotations. This class is designed to be domain-neutral, adapting various raw data structures. Each specific dataset, with its unique file structure and domain, inherits from the review class and customizes the loader function. Consequently, the output is a standardized review object. This uniformity allows various aspect detection and sentiment analysis models to operate with the review object, eliminating the need for models to adapt to each dataset's unique structure and format. After transforming the reviews within the dataset into the standardized review object format, they become prepared for input into the aspect model layer (aml) for model training purposes. However, before moving forward to the aspect detection phase and in line with this focus on empowering the models for latent occurrences of aspects, the processed reviews will be first passed to the data augmentation layer.

3.2 Data Augmentation

Upon converting reviews from the dataset into a standardized review object format, these reviews are directed to an augmentation layer. This layer is specifically designed to improve the performance of aspect detection models. Our research identifies inherent limitations in existing aspect detection models, notably in aspect term and category detection. Aspect term detection models, for example, falter when they encounter novel terms or phrases not present in their training data. This limitation becomes apparent in scenarios where new concepts emerge and are not recognized by the models. Aspect category detection models, conversely, are constrained by their dependence on a fixed set of categories. They might broadly

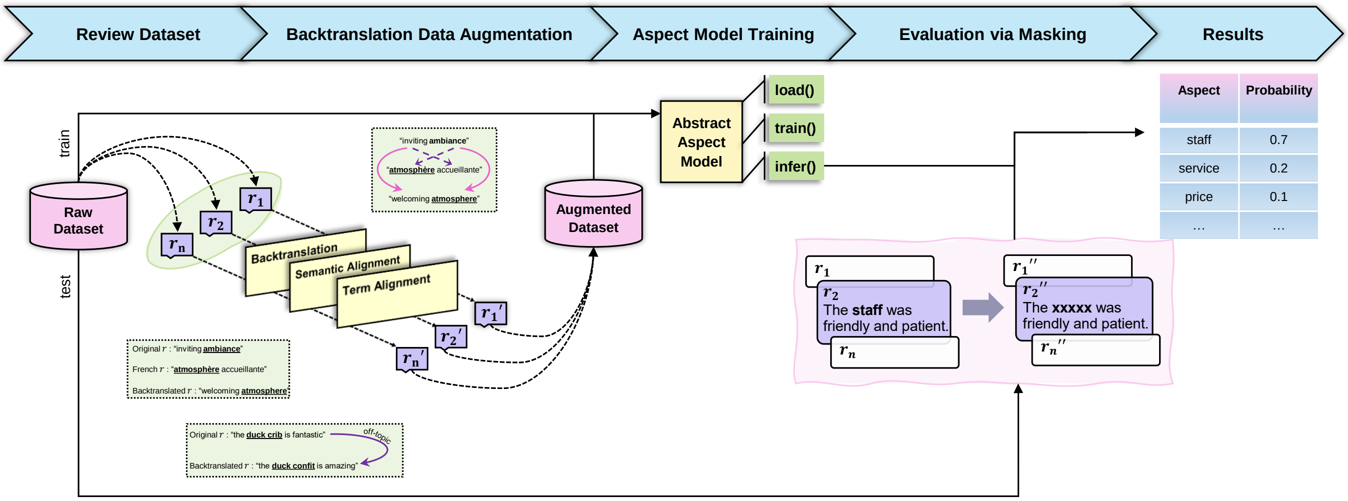

Figure 1: Comprehensive workflow diagram of LADy toolkit: from data input to outcome results.

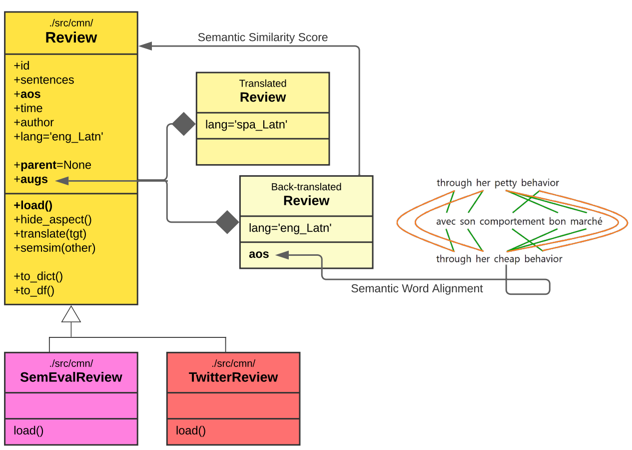

Figure 2: Illustration of the review class.

```python
# ./src/cmn/review.py
class Review(object):
    def __init__(self, id, sentences, time, author, aos, parent,
        lang):
        self.id = id
        self.sentences = sentences
        self.time = time
        self.author = author
        self.aos = aos
        self.lang = lang
        self.parent = parent
        self.augs = {}
    def hide_aspects(self, mask):
    def translate(self, tgt, settings): ...
    def semsim(self, other):
    def semalign(self, other): ...
    def preprocess(self): ...
    def load(path): ...
    def get_aos(self): ...
    def get_txt(self): ...
    def get_stats(datapath, output): ...
    def plot_dist(stats, output): ...
```

Figure 3: Review class attributes and methods.

categorize a review under *'food'* but lack the granularity to distinguish between sub-categories like desserts or specific dishes. Moreover, these models typically rely on explicit aspects that are directly mentioned in reviews.

Despite these challenges, enhancing existing models is feasible without the need for designing new aspect detection methodologies and can be accomplished through dataset augmentation in the training phase. In this work, we adopt backtranslation augmentation to enrich review datasets with a wider range of vocabulary. Backtranslation involves translating a sentence to a different language and then translating it back to the original language, while preserving its meaning. This method adds contextually relevant synonymous terms to the original language, thereby enriching the training datasets. For instance, it might replace *'food'* with *'dish,'* adding depth to the contextual understanding. Utilizing machine translation, we experiment with translating reviews into a different language (e.g., `French`) and then back to `English`, exploring the potential of backtranslation for semantic augmentation. This process helps in uncovering latent aspects, addressing out-of-vocabulary terms, and clarifying ambiguous expressions.

For our translations, we employ Meta's *'no language left behind'* (`nllb`) [4], a versatile neural machine translator supporting 200 languages, chosen for its emphasis on universal translation and its capability to serve low-resource language communities. Post-backtranslation, each review, now appended with a language attribute, is re-incorporated into the review dataset for subsequent training phases, as illustrated in Figure 1. However, backtranslation presents its own challenges. There is a risk of topic drift in content during translation, which may lead to the loss of explicit aspect labels. To mitigate this, we have implemented a control mechanism that involves reviewing the backtranslated content and its corresponding labels. This process is depicted in Figure 1, forming a part of our primary workflow for augmentation, which consists of two main steps:

Step 1: Semantic Alignment

We employ DeCLUTR [6] for ensuring the semantic consistency. This tool compares the original and backtranslated reviews by assigning a semantic score. Reviews scoring below 0.5 suggest the topic drift and will be excluded from the augmented dataset, and this score will be recorded in the `semsim` attribute of the augmented review object.

Step 2: Term Alignment
Recognizing that backtranslation may alter the wording, we utilize `SimAlign` [10] to ensure accurate labeling. `SimAlign` provides index mappings for word alignments between paraphrased or translated sentences, aiding in identifying new aos for the augmented reviews, which are then stored in their respective review objects.

Through these processes, we aim to enhance the aspect detection capability of our models. A key point to underscore is that LADy is designed to use any augmentation technique. This flexibility is achieved by adding each augmented review as an attribute to the original review object, regardless of the augmentation method utilized, whether it's backtranslation, as demonstrated in our example, or any other method. It is also important to note that the augmentation phase is optional, and we can proceed directly to the aspect detection phase and train the models using the original datasets. Having navigated beyond the augmentation layer, reviews are now prepared for the training phase.

3.3 Aspect Model

The aspect model layer (aml) features the abstract class AbstractAspectModel, specifically customized for aspect modeling techniques. Figure 5 displays its key methods and attributes, offering a clear understanding of its structure and functionality. At the core of aml is the AbstractAspectModel, which serves as a standardized interface for aspect detection. This abstract class is a crucial component of LADy, allowing for the integration and management of various aspect detection models. Each model added to LADy as an aspect detection model is hosted by this class and is required to override specific functions, such as training and inference, as defined by the abstract class. This ensures that all models conform to a uniform operational toolkit. Moreover, the AbstractAspectModel encompasses functions that are universally applicable across different models, such as evaluation and preprocessing. These functions are implemented directly in the abstract class, providing a standardized approach for fair preprocessing and evaluation. This uniformity is important for maintaining consistency and fairness in processing and assessing the performance of diverse aspect detection models. The use of an abstract class like the AbstractAspectModel, allows for the creation of a template that defines a standard set of behaviors and properties, which can then be inherited and customized by various emerging aspect detection models.

Initially, the model undergoes training using input training and validation data, and the resulting model is stored in the output. LADy incorporates a combination of traditional topic models and neural topic models as well as aspect detection models, both in unsupervised and supervised settings, as baselines. Figure 5 shows the class hierarchy within these aspect detection methods. For training and comparative analysis against other baselines, the models are divided into two categories. The first category focuses on baselines

```python
# ./src/aml/mdl.py
class AbstractAspectModel:
    def __init__(self, naspects, nwords):
        self.naspects = naspects
        self.nwords = nwords
        self.dict = None
    def load(path): ...
    def train(reviews_train, reviews_valid, settings, output): ...
    def quality(metric): ...
    def infer(reviews, doctype): ...
    def preprocess(reviews, settings=None): ...
```

Figure 4: AbstractAspectModel class in aspect model layer.

that rely on topic modeling methods explicitly used for detecting topics within documents. Each review is considered a separate document, with an associated distribution of topics, thus uncovering the most probable words indicative of the latent or explicit aspects of the review. The models utilized in this category include:

- **`loclda`** [2]: A traditional topic model that identifies aspects by associating highly probable words with topics.
- **`btm`** [12]: Another traditional topic model that detects aspects based on patterns of word co-occurrence.
- **`ctm`** [1]: An unsupervised aspect detection model that combines neural networks with contextual topic modeling.
- **`neurallda`** [21]: An unsupervised aspect detection model that integrates neural networks and topic modeling.

On the other hand, the second category comprises methodologies primarily designed for the specific task of aspect detection. The models employed in this category include:

- **`random`**: This technique leverages a random selection process, extracting aspects from the list of words present within the training set.
- **`cat`** [23]: An unsupervised aspect detection that classifies the review with predefined aspects.
- **`bert-tfm`** [14]: A supervised sequential labeling model that harnesses the power of BERT, integrating a straightforward linear classification layer to streamline the aspect detection process.

Moreover, the toolkit has an aspect detection interface designed to facilitate the easy customization of training and inference procedures for each newly integrated model. These models function effectively as black boxes, requiring only specific datasets and hyperparameter settings to produce outputs, which include aspects and their corresponding evaluations. Subsequently, we employ this model to make inferences (predictions) regarding the aspect of a given review as shown in Figure 4. Up next, we will explore the adaptable and reproducible evaluation methodology implemented. Special attention will be given to managing the underlying occurrences of latent aspects.

3.4 Evaluation

The evaluation phase of LADy showcases its innovative approach to handling latent datasets for aspect detection. While our goal was to establish a toolkit for detecting latent aspects, we found that, to the best of our knowledge, there are no existing public datasets

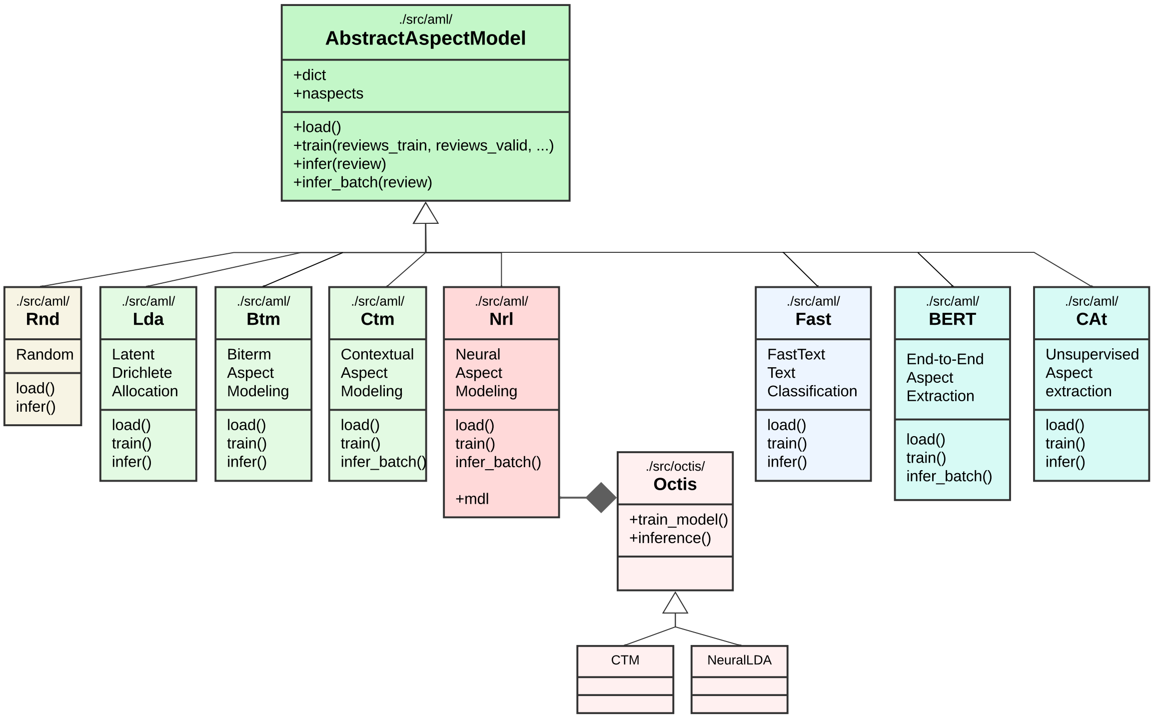

Figure 5: Class hierarchy in the aspect model layer.

featuring reviews with latent aspect labels that lack explicit representation in the text. Given the absence of pre-labeled datasets with latent aspects, LADy defines a masking method that masks a specific proportion of explicit aspect labels within reviews to mimic the latency. Notably, this masking is not mandatory, and users have the option to bypass it by setting the masking proportion (*h_ratio*) to 0.0. This attribute, ranging from 0.0 (no latent aspects meaning that all aspects are explicit) to 1.0 (all aspects are latent), is adjustable in the hyperparameters' setting shown in Figure 6, offering flexibility in dataset creation. This strategy cultivates datasets rich in latent aspects, crucial for assessing the system's proficiency in the detection of latent aspects.

Later after masking the test datasets during the evaluation phase, the system generates a likely set of words to represent aspects and assesses these aspects using a variety of metrics, benchmarking them against a predefined reference or 'golden set.' The comprehensive list of evaluation metrics includes, but is not limited to precision (`p`), recall (`r`), normalized discounted cumulative gain (`ndcg`), mean average precision (`map`), and success (`s`). Additionally, these metrics are provided with values @k, indicating the model's performance across various levels of probable aspect lists. This feature is particularly relevant for latent aspect detection, where a single aspect may not be sufficient, and a ranked list of aspects offers greater insights. The toolkit's flexibility extends to the evaluation of explicit aspects, allowing for the specification of k in the parameters (also accessible in Figure 6) that mostly is 1 in the case of explicit aspect detection. This adaptability is not limited to the value of k alone but also includes the easy integration of new information retrieval evaluation metrics by simply adding these metrics to the toolkit parameters file.

```
# ./src/params.py
settings = {
    'cmd': ['prep', 'train', 'test', 'eval', 'agg']
    'prep': {'langaug': 'pes_Arab', 'spa_Latn', ...]},
    'train': {'ratio': 0.85, 'nfolds': 5, 'nwords': 20,
              'langaug_semsim': 0.5,
              'qualities': ['coherence', 'perplexity']},
    'test': {'h_ratio': 0.5},
    'eval': {'metrics': ['precision', 'recall', ...],
             'topkstr': [1, 5, 10]}
```

Figure 6: The hyperparameters of LADy for each layer.

4 SYSTEM USAGE

The LADy toolkit is designed to accommodate a diverse user base, from experienced `python` developers to those preferring a more straightforward, user-friendly interface. For advanced users, LADy functions as a sophisticated `python` library, offering deep customization options and extensive control over its various aspect models and parameters. This allows experienced programmers and data scientists to fully exploit its capabilities in a programmable environment. Simultaneously, LADy features an intuitive web dashboard, accessible to users of all skill levels. This dashboard simplifies the user experience by providing a clear, navigable interface for experimenting with models, setting parameters, and viewing results.

Table 2: Statistics of the datasets: showcasing dataset sizes and average aspect counts per review.

dataset	#review	avg #aspects
semeval-14-laptop	1,488	1.5846
semeval-14-restaurant	2,023	1.8284
semeval-15-restaurant	833	1.5354
semeval-16-restaurant	1,234	1.5235
twitter	6,248	1.0000

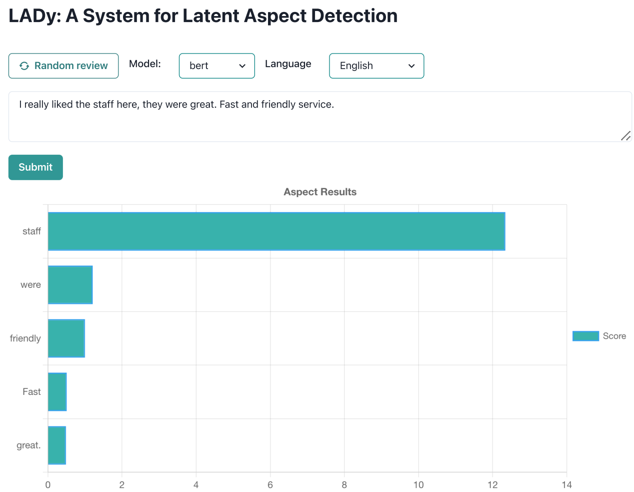

Figure 7: Evaluation and aspect scores visualizing the evaluation metrics for aspect detection models

4.1 Quick Start

To quickly start using LADy, you can obtain it by cloning the GitHub repository at: `github.com/fani-lab/LADy`. LADy incorporates various aspect models that have specific hyperparameters, which can be set in the `./src/params.py` file and presented in Figure 6 that provides a summary of these hyperparameters. Before training the model, you can adjust the number of aspects (`#naspects`) to suit your requirements. Additionally, LADy offers a host of augmentation methods and specifically utilizes a backtranslation method during the training phase to enhance the results. You have the option to provide a list of languages for the backtranslation process on the training dataset. The entry point to the toolkit is `./src/main.py`, which executes the entire pipeline, leading to the delivery of aspects for each review, followed by evaluation metrics of top k results. For a quick start experience, a toy sample of reviews is provided as well.

4.2 Case Study

In our study, we deployed the LADy toolkit across various versions and domains of the `semeval` datasets [16–18], encompassing both the restaurant and laptop domains, as well as `twitter` dataset [5] in social media domains. Table 2 demonstrates the statistics about the datasets currently available.

This application was aimed at evaluating and comparing different aspect detection methodologies in a consistent toolkit and experimental results rely heavily on the selection of baseline models rather than on the toolkit introduced in this paper. However, as illustrated in Table 3, the state-of-the-art methods demonstrated considerable proficiency in identifying both latent and explicit aspects in all the languages and specifically in the case where we had the combination of all languages together as one dataset called *'all'*. This was quantified using a range of information retrieval evaluation metrics. A notable enhancement in model performance was observed when augmentation techniques were employed during the training phase, strengthening the models' capability to detect the latent and explicit aspects. Comprehensive results, including those from other datasets, baseline comparisons, and additional metrics, are available in the LADy's codebase repository.

4.3 Web Application

This section delves into our web application, emphasizing its versatile parameterization feature for aspect detection. The application serves as a dynamic platform accommodating various aspect detection models, allowing researchers to experiment with methodologies on selected input reviews. The user-friendly interface facilitates interaction with different aspect detection methods, particularly highlighting latent aspect detection and the impactful integration of backtranslation during the models' training phase. Researchers can leverage the model comparison functionality within the application, allowing for a direct comparison between models with and without backtranslation. Figure 7 presents the evaluation and aspect list scores for one of the baselines which is `bert` representing the model trained on the original dataset for the case of explicit occurrence of an aspect.

5 CONCLUSIONS

This paper introduced LADy, a robust toolkit designed for conducting reproducible research in aspect-based sentiment analysis, with a particular emphasis on latent aspects. LADy incorporates various aspect detection models, and diverse datasets for review analysis, including unsolicited reviews, and employs multiple evaluation strategies. It also offers a feature augmentation option during the training phase, all within a flexible and adaptable integrated structure. Our future work will involve expanding LADy's capabilities to encompass sentiments and opinions, allowing its ability to handle a wider spectrum of sentiment analysis tasks.

REFERENCES

[1] Federico Bianchi, Silvia Terragni, and Dirk Hovy. 2021. Pre-training is a Hot Topic: Contextualized Document Embeddings Improve Topic Coherence. 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)*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, Online, 759–766. https://doi.org/10.18653/v1/2021.acl-short.96

[2] Samuel Brody and Noemie Elhadad. 2010. An Unsupervised Aspect-Sentiment Model for Online Reviews. In *NAACL 2010*. 804–812. https://aclanthology.org/N10-1122/

[3] Hao Chen, Zepeng Zhai, Fangxiang Feng, Ruifan Li, and Xiaojie Wang. 2022. Enhanced Multi-Channel Graph Convolutional Network for Aspect Sentiment Triplet Extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational

Table 3: The average performance of some of the aspect detection baselines with backtranslation augmentation evaluated on one of the datasets. The highest performing results are highlighted in bold, and the second best are underlined.

	bert-tfm [14]			cat [23]			loclda [2]			btm [12, 29]			ctm [1]			random		
	pr1	rec5	ndcg5	pr1	rec5	ndcg5	pr1	rec5	ndcg5	pr1	rec5	ndcg5	pr1	rec5	ndcg5	pr1	rec5	ndcg5
	semeval-15-restaurant																	
none	**0.7000**	0.6897	0.6757	0.3327	0.5248	0.4343	0.2320	0.3549	0.2925	0.1872	**0.3133**	0.2500	0.0560	0.0493	0.0485	0.0000	0.0005	0.0005
+Chinese	0.6661	0.6928	0.6699	0.3723	0.5287	0.4596	0.1968	0.3408	0.2647	0.1760	0.2783	0.2261	0.0624	0.0717	0.0637	**0.0016**	0.0028	0.0022
+Farsi	0.6742	0.6707	0.6608	0.3703	0.5386	0.4592	0.1840	0.3494	0.2689	0.1776	0.2834	0.2303	0.0560	0.0823	0.0722	0.0000	0.0002	0.0002
+Arabic	0.6661	0.6898	0.6671	0.4139	0.5683	0.4939	0.2000	0.3654	0.2887	0.1568	0.2956	0.2269	0.0592	0.0649	0.0577	0.0000	0.0000	0.0000
+French	0.6565	0.7030	0.6734	0.4040	0.5584	0.4883	0.2512	0.3577	0.3032	0.1968	0.3048	0.2481	0.0720	0.0837	0.0733	0.0000	0.0008	0.0006
+German	0.6710	0.6927	0.6721	0.3980	0.5505	0.4787	0.2416	0.3648	0.2976	0.1808	0.2691	0.2242	0.0560	0.0717	0.0603	0.0000	**0.0061**	**0.0036**
+Spanish	0.6645	0.7099	0.6769	0.3921	0.5663	0.4842	0.2224	0.3737	0.3035	0.2000	0.2975	0.2466	0.0464	0.0531	0.0458	0.0000	0.0000	0.0000
+all	0.6613	**0.7182**	**0.6823**	**0.5980**	**0.7861**	**0.7096**	**0.2592**	**0.3744**	**0.3104**	**0.2128**	0.2986	**0.2515**	**0.2192**	**0.2470**	**0.2263**	**0.0016**	0.0008	0.0010

Linguistics, Dublin, Ireland, 2974–2985. https://doi.org/10.18653/v1/2022.acl-long.212

[4] Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, and et al. 2022. No Language Left Behind: Scaling Human-Centered Machine Translation. *CoRR* abs/2207.04672 (2022). arXiv:2207.04672

[5] 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 (Volume 2: Short Papers)*, Kristina Toutanova and Hua Wu (Eds.). Association for Computational Linguistics, Baltimore, Maryland, 49–54. https://doi.org/10.3115/v1/P14-2009

[6] 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)*. Association for Computational Linguistics, Online, 879–895. https://doi.org/10.18653/v1/2021.acl-long.72

[7] Zhibin Gou, Qingyan Guo, and Yujiu Yang. 2023. MvP: Multi-view Prompting Improves Aspect Sentiment Tuple Prediction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 4380–4397. https://doi.org/10.18653/v1/2023.acl-long.240

[8] Farinam Hemmatizadeh, Christine Wong, Alice Yu, and Hossein Fani. 2023. Latent Aspect Detection via Backtranslation Augmentation. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management* (Birmingham, United Kingdom) *(CIKM '23)*. Association for Computing Machinery, New York, NY, USA, 3943–3947. https://doi.org/10.1145/3583780.3615205

[9] Ehsan Hosseini-Asl, Wenhao Liu, and Caiming Xiong. 2022. A Generative Language Model for Few-shot Aspect-Based Sentiment Analysis. In *Findings of the Association for Computational Linguistics: NAACL 2022*, Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (Eds.). Association for Computational Linguistics, Seattle, United States, 770–787. https://doi.org/10.18653/v1/2022.findings-naacl.58

[10] Masoud Jalili Sabet, Philipp Dufter, François Yvon, and Hinrich Schütze. 2020. SimAlign: High Quality Word Alignments Without Parallel Training Data Using Static and Contextualized Embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, Online, 1627–1643. https://doi.org/10.18653/v1/2020.findings-emnlp.147

[11] Baoxing Jiang, Shehui Liang, Peiyu Liu, Kaifang Dong, and Hongye Li. 2023. A semantically enhanced dual encoder for aspect sentiment triplet extraction. *Neurocomputing* 562 (2023), 126917. https://doi.org/10.1016/j.neucom.2023.126917

[12] Ning Li, Chi-Yin Chow, and Jia-Dong Zhang. 2019. Seeded-BTM: Enabling Biterm Topic Model with Seeds for Product Aspect Mining. In *21st IEEE International Conference on High Performance Computing and Communications*. IEEE, 2751–2758.

[13] 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-18*. International Joint Conferences on Artificial Intelligence Organization, 4194–4200. https://doi.org/10.24963/ijcai.2018/583

[14] Xin Li, Lidong Bing, Wenxuan Zhang, and Wai Lam. 2019. Exploiting BERT for End-to-End Aspect-based Sentiment Analysis. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*. Association for Computational Linguistics, Hong Kong, China, 34–41.

[15] Joseph Peper and Lu Wang. 2022. Generative Aspect-Based Sentiment Analysis with Contrastive Learning and Expressive Structure. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (Eds.). Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 6089–6095. https://doi.org/10.18653/v1/2022.findings-emnlp.451

[16] Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad AL-Smadi, and et al. 2016. SemEval-2016 Task 5: Aspect Based Sentiment Analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, Steven Bethard, Marine Carpuat, Daniel Cer, David Jurgens, Preslav Nakov, and Torsten Zesch (Eds.). Association for Computational Linguistics, San Diego, California, 19–30. https://doi.org/10.18653/v1/S16-1002

[17] Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Suresh Manandhar, and Ion Androutsopoulos. 2015. SemEval-2015 Task 12: Aspect Based Sentiment Analysis. In *Proceedings of the 9th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2015*. The Association for Computer Linguistics, 486–495.

[18] 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@COLING*. The Association for Computer Linguistics, 27–35.

[19] Tian Shi, Liuqing Li, Ping Wang, and Chandan K Reddy. 2021. A simple and effective self-supervised contrastive learning framework for aspect detection. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 35. 13815–13824.

[20] Youwei Song, Jiahai Wang, Tao Jiang, Zhiyue Liu, and Yanghui Rao. 2019. Targeted sentiment classification with attentional encoder network. In *Artificial Neural Networks and Machine Learning–ICANN 2019: Text and Time Series: 28th International Conference on Artificial Neural Networks, Munich, Germany, September 17–19, 2019, Proceedings, Part IV 28*. Springer, 93–103.

[21] Akash Srivastava and Charles Sutton. 2017. Autoencoding Variational Inference For Topic Models. In *5th International Conference on Learning Representations, ICLR 2017*. OpenReview.net. https://openreview.net/forum?id=BybtVK9lg

[22] Mohammad Tubishat, Norisma Idris, and Mohammad Abushariah. 2021. Explicit aspects extraction in sentiment analysis using optimal rules combination. *Future Generation Computer Systems* 114 (2021), 448–480. https://doi.org/10.1016/j.future.2020.08.019

[23] Stéphan Tulkens and Andreas van Cranenburgh. 2020. Embarrassingly Simple Unsupervised Aspect Extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (Eds.). Association for Computational Linguistics, Online, 3182–3187. https://doi.org/10.18653/v1/2020.acl-main.290

[24] Manju Venugopalan and Deepa Gupta. 2022. An enhanced guided LDA model augmented with BERT based semantic strength for aspect term extraction in sentiment analysis. *Knowledge-Based Systems* 246 (2022), 108668. https://doi.org/10.1016/j.knosys.2022.108668

[25] Manju Venugopalan and Deepa Gupta. 2022. A reinforced active learning approach for optimal sampling in aspect term extraction for sentiment analysis. *Expert Systems with Applications* 209 (2022), 118228. https://doi.org/10.1016/j.eswa.2022.118228

[26] An Wang, Junfeng Jiang, Youmi Ma, Ao Liu, and Naoaki Okazaki. 2023. Generative Data Augmentation for Aspect Sentiment Quad Prediction. In *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)*, Alexis Palmer and Jose Camacho-collados (Eds.). Association for Computational Linguistics, Toronto, Canada, 128–140. https://doi.org/10.18653/v1/2023.starsem-1.12

[27] Qianlong Wang, Zhiyuan Wen, Qin Zhao, Min Yang, and Ruifeng Xu. 2021. Progressive Self-Training with Discriminator for Aspect Term Extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language*

Processing, Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (Eds.). Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 257–268. https://doi.org/10.18653/v1/2021.emnlp-main.23

[28] Zengzhi Wang, Qiming Xie, and Rui Xia. 2023. A Simple yet Effective Framework for Few-Shot Aspect-Based Sentiment Analysis. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1765–1770.

[29] Xiaohui Yan, Jiafeng Guo, Yanyan Lan, and Xueqi Cheng. 2013. A biterm topic model for short texts. In *22nd International World Wide Web Conference, WWW '13, Rio de Janeiro, Brazil, May 13-17, 2013*. International World Wide Web Conferences Steering Committee / ACM, 1445–1456.

[30] Heng Yang, Chen Zhang, and Ke Li. 2022. PyABSA: A Modularized Framework for Reproducible Aspect-based Sentiment Analysis. *arXiv preprint arXiv:2208.01368* (2022).

[31] Yunyi Yang, Kun Li, Xiaojun Quan, Weizhou Shen, and Qinliang Su. 2020. Constituency Lattice Encoding for Aspect Term Extraction. In *Proceedings of the 28th International Conference on Computational Linguistics*. International Committee on Computational Linguistics, Barcelona, Spain (Online), 844–855. https://doi.org/10.18653/v1/2020.coling-main.73

[32] Guoxin Yu, Jiwei Li, Ling Luo, Yuxian Meng, Xiang Ao, and Qing He. 2021. Self Question-answering: Aspect-based Sentiment Analysis by Role Flipped Machine Reading Comprehension. In *Findings of the Association for Computational Linguistics: EMNLP 2021*. Association for Computational Linguistics, Punta Cana, Dominican Republic, 1331–1342. https://doi.org/10.18653/v1/2021.findings-emnlp.115