E-ISSN 2224-2333
P-ISSN 2222-9930

Pakistan Journal of Engineering Technology and Science (PJETS)
Volume 13. Issue.2, PP.116-124, December 2025
https://doi.org/10.22555/pjets.v13i2.1417

HS-ABSA: An Annotated Dataset for Aspect-Based Sentiment Analysis of Home Services Customer Reviews

Afsheen Maroof [1], Muhammad Hussain Mughal[2,*], Shaukat Wasi[3]
[1] *Department of Computer Science, Air University Karachi*
[2] *Department of Computer Science, Sukkur IBA University, Sukkur, Pakistan*
[3] *Department of Computer Science, Muhammad Ali Jinnah University, Karachi, Pakistan*
[*] *Corresponding author: muhammad.hussain@iba-suk.edu.pk*

Abstract:

With the progression of new technologies, the explosive growth of user-generated content on online platforms has significantly fuelled interest in Aspect-Based Sentiment Analysis (ABSA) for both academic research and commercial applications. While existing ABSA research often utilizes publicly available product-oriented datasets such as those for laptops and cameras, there remains a scarcity of high-quality, publicly available datasets tailored for service-oriented domains. This gap highlights the need to develop new domain-specific resources that can effectively support the evaluation of ABSA models in real-world customer service contexts. In this paper, we present HS-ABSA, a new benchmark human-annotated dataset specifically curated for aspect-based sentiment analysis of home services customer feedback. The dataset was constructed by scraping reviews from the Google Play Store for a mobile app. The scraped reviews were carefully analysed, filtered, and manually annotated. The annotation process focused on opinion terms, aspect categories, and sentiment polarity. Inter-annotator agreement was measured using Kappa's score. A total of 4,164 reviews were manually annotated. The inter-annotator agreement was calculated using Kohen's kappa, and scores of 91.92% for opinion terms, 84.00% for aspect categories, and 94.38% for sentiment polarity were obtained. This newly curated ABSA Dataset for the service industry will provide a robust foundation for evaluating ABSA models.

Keywords: *Dataset, Home Service Reviews, Aspect-Based Sentiment Analysis, Sentiment Analysis, Opinion Term, Aspect Category, Sentiment Polarity*

I. INTRODUCTION

The dramatic increase in web technologies and the rapid increase in smartphone apps have helped users to share their opinions and thoughts on different social media platforms. Sentiment Analysis is the task of identifying a user's feelings, opinions, and thoughts about a specific product or service. Sentiment analysis plays a crucial role for organizations aiming to gauge the public perception of their products and services [1]. Sentiment analysis at the aspect level is a detail-level analysis where opinions consist of targets, and the sentiments associated with each target are determined [2].

Existing research shows that most of the publicly available datasets used for aspect-based sentiment analysis (ABSA) focus on products like mobiles, cameras, and restaurants, making domain-specific datasets challenging to find [3], [4] [5] for example, [6] used the electronic dataset to extract the features of a product from customer reviews, similarly [7] used a movie review dataset to extract the sentiments attached to different aspects of a

movie. AWARE is a newly created dataset for ABSA by [8], in which they collected 11323 reviews from three major domains of MobileApp. One of the most popular shared tasks on ABSA is the SemEval 2014 task 4 by [9], which attracted a large number of teams and submissions. They published a dataset on ABSA, annotated with both term-level and document-level labels. They annotated reviews of laptops and restaurants. Several ABSA datasets in eight different languages were released [10]; it was also a follow-up of SemEval 2014, where text-level aspect annotations were also introduced. SamEval workshops are considered a major contribution to providing the datasets for ABSA; however, datasets constructed in these workshops are also product-oriented. However, finding a domain-specific service-oriented dataset publicly is challenging. Service reviews differ from product reviews. Rather than evaluating physical features such as battery life, screen resolution, or build quality (as is common in product reviews), service review is often using soften, subjective, context-sensitive, and loaded with implicit sentiments. Users in the service domain describe dynamic experiences such as punctuality, professionalism, and staff behaviour. This lack of domain-specific resources restricts progress in developing and evaluating ABSA models suited to service contexts. Therefore, there arises a need to construct a data set that should cater to the ABSA tasks for the service industry. This dataset addresses this gap by providing a specialized resource for the home service industry, allowing researchers to explore sentiment trends, customer preferences, and expectations within this domain. The dataset includes implicit targets and features commonly found in service industry reviews, enhancing its real-world applicability.

II. RELATED STUDY

Researchers have discussed various methods to construct dataset resources in the field of ABSA. [9] Presented the ABSA task in SemEval-2014 [9]. This task attracted numerous teams and submissions, contributing significantly to the field. The organizers published a comprehensive dataset annotated with term- and document-level labels, focusing on reviews of laptops and restaurants. Table 1 shows the details of the ABSA dataset and its annotated tasks.

Table 1: ABSA Dataset and its Annotated Tasks

Reference	Domain	Task Annotated	# of reviews
[9]	Laptop reviews Restaurant reviews:	Aspect Term Aspect Polarity Aspect category	3,845 sentences 3,041 sentences
[11]	Selected Book Reviews	Aspect Term Extraction, Aspect Term Polarity Aspect Category selection	1,513
[14]	Restaurants Laptops Hotels shopping reviews	Aspect of Category Extraction Aspect of Term Extraction Sentiment Polarity Detection	2,250 3,027 2,904
[15]	Location-Based ABSA	Aspect Category, Aspect Polarity	3862
[16]	Restaurants	Aspect Term, Aspect Polarity	8,879
[17]	App reviews	Aspect categories, Aspect Polarities	3,148

Al-Smadi et al. provided a benchmark dataset, HAAD[1], of Arabic text for ABSA by annotating 1,513 selected book reviews [11]. This dataset comprises four major ABSA tasks: Aspect Term Extraction (ATE), Aspect Term Polarity Identification, Aspect Category Detection (ACD), and Aspect Category Polarity Identification.
Following the 2014 task, Pontiki et al. in [10] continued the momentum and organized SemEval-2016 Task 5 and released ABSA datasets in eight different languages. The English dataset in this task comprises 900 reviews (5,801 sentences) related to restaurants and laptops. In contrast to SemEval-2014, a key innovation in SemEval-2016 was

[1] Human annotated Arabic dataset

the introduction of text-level aspect annotations, where aspect labels are assigned to entire reviews instead of individual sentences.

Saeidi et al. developed the SentiHood dataset for targeted aspect-based sentiment analysis, using location as the focal entity and constraining their study to neighbourhoods in the city of London (Saeidi et al., 2016). They annotated only reviews that mentioned at least two locations. A challenging dataset known as the MAMS dataset was constructed by Jiang et al. (2019) for the ABSA task. They annotated 8,879 reviews, ensuring that each review contained at least two different opinions with distinct sentiment polarities.

Alqaryouti et al. proposed a new dataset for government apps of Dubai city, consisting of mobile app opinions and associated aspects [12]. They use a manual annotation process to construct the data set. Annotators used GARSA to annotate reviews, ensuring comprehensive and accurate data labelling. A benchmark dataset was proposed [13] in the Urdu language. They annotated four tasks of ABSA AWARE, a newly created dataset for ABSA by Alturaief et al. in [8]. They collected 11,323 reviews from three significant domains of smartphone apps, providing a comprehensive resource for ABSA. In their work, Kontonatsios et al. introduced FABSA, a large-scale dataset spanning 10 different domains for extraction [14].

III. MATERIALS AND METHODS

The ABSA dataset for the service industry aims to pinpoint the individual opinion terms or targets that are expressed within the review, assign a general aspect category to each extracted term, and evaluate the corresponding sentiment orientation of each aspect present in the review text. Sentiment orientations are mapped to positive and negative depending on the words or phrases that describe each aspect. The framework is illustrated in Figure 1. It is composed of three key steps, namely, data collection, data filtration, and data annotation.

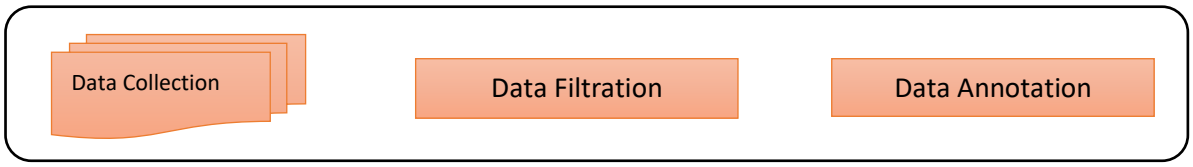

Figure 1: Dataset Construction Framework

A. Data Collection

The primary objective of this research is to construct a dataset for a home service company. To achieve this, reviews of the company's mobile application were scraped from the Google Play Store. In particular, the study focuses on UrbanClap (UC), a home maintenance platform that provides reliable services with an emphasis on customer comfort and safety.

The data collection process was carried out using the Python library google-play-scraper[2].The extracted dataset includes the following attributes: username, review content, review timestamp, rating score, and thumbs-up count. In total, 131,655 reviews were collected. The majority of the reviews had high ratings. For instance, 76,450 reviews were rated 5 stars, 9,072 reviews received 4 stars, 3,488 reviews received 3 stars, 3,697 reviews received 2 stars, and 38,948 reviews were rated 1 star. In this study, only reviews written in English were considered, as they represented the majority of the collected data.

B. Data Filtration

The dataset was further refined through a series of pre-processing steps. An initial analysis revealed that emoticons were rarely used in the app reviews; therefore, reviews containing emoticons were excluded from the dataset. Similarly, single-word reviews were removed, as they do not provide meaningful insights. Reviews consisting of incomplete or purely subjective sentences that lacked opinion-bearing terms were also omitted.

[2] https://pypi.org/project/google-play-scraper/

As the data originated from real-world sources, many reviews contained grammatical and spelling errors. These issues were systematically corrected during the pre-processing stage. After applying all filtration steps, the dataset was reduced to 4,164 reviews, which were then used for subsequent phases of analysis. Filters applied in pre-processing step are shown in Figure 2.

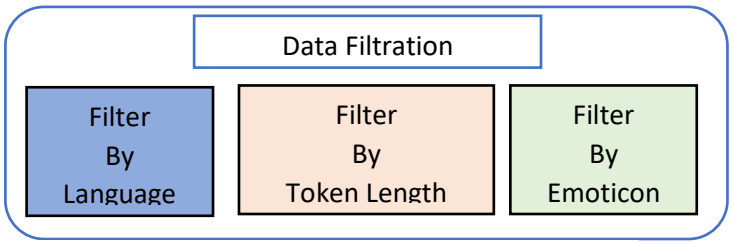

Figure 2: Overview of Data Filtration Step

C. Data Annotation

Filtered data is then passed to the annotation process. Figure 3 depicts the complete annotation process. A manual annotation process is opted for this purpose, although the manual annotation process is time-consuming and needs more human resources, but it is more effective and gives better results [12].

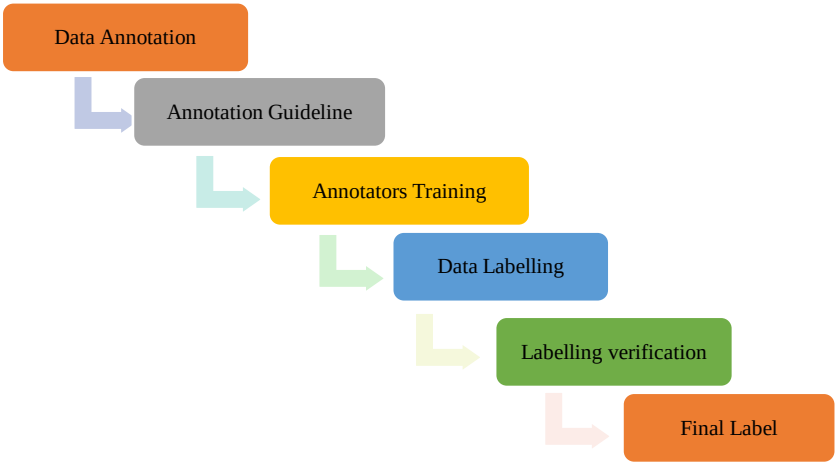

Figure 3: Data Annotation pipeline for home service app reviews

D. Annotation Guidelines

Before commencing with the manual annotation process, some general guidelines were prepared for annotating the data. These guidelines were prepared following the SemEval workshop [3] and the quality guidelines provided by the Google Play Store[4] to develop a mobile app to identify the terms and sentiment polarity. The following three types of information must be extracted and annotated from the given reviews.

- Opinion Term/Terms
- Aspect Category
- Aspect Polarity

Opinion Term Guidelines:

- A noun/noun with an adjective expressing an opinion will be extracted as a term.
- If there is more than one term in a sentence, then annotate all of them.
- If the identified term is misspelt, it should be annotated after spelling correction

Aspect Categories Guidelines: The most used smartphone operating system is Android (by Google Inc.), which boasts a global market share of 86.2%. Android provides a set of quality guidelines to be considered while

[3] https://alt.qcri.org/semeval2016/task5/data/uploads/absa2016_annotationguidelines.pdf
[4] https://developer.android.com/docs/quality-guidelines/core-app-quality

developing an Android mobile app. Also, certain aspects must be considered while looking at the customer perspective of a service-providing company. Aspect categories were designed according to the provided standard of SemEval and the Android app from the customer's perspective. Based on these two factors, we defined 10 general categories. The complete description is presented below.

Application UI: This category is related to the design of the application. Design should be simple, user-friendly, appealing, and attractive.
Application Functionality This category is concerned about users' experience and their level of satisfaction with the functionality of the app, like how the app works, whether it is easy to avail a service using the app, can it easily navigating is it syncs with new updates.
App Utility: This category is concerned with the usage of the application. Is the app good or bad?
Company Service This category is related to the overall experience of the customer and their satisfaction with the service of the company.
Service Quality This category includes the experience of users related to any specific service they received from the
Service person's attitude. This aspect deals with the behaviour of the person who was assigned by the company to complete the service.
Customer Support: This category is concerned with what customer thinks about customer support services, like how supportive the representative is? And how quickly they resolve the issues of customers?
Cost: All the expenses related to this category

Service schedule: This category is concerned with the response to customer requests. When the request for a service is made by a customer, it should be scheduled according to the customer.
Payment: The payment-related issues, like advance payments, not refunding, delayed payments, etc., are included.

Aspect Polarities Guidelines:

- Assign a positive or negative polarity to each extracted aspect term.
- If a sentence contains positive sentiment, assign a positive polarity to it.
- If a sentence contains negative sentiment, assign a negative polarity to it.

E. Manual Annotation

The annotation process mainly comprises 3 sub processes, namely, annotation, verification, and approval. A team of four members was involved in the whole process. Two members were master's students of the Natural Language Processing research project, and they were asked to annotate the data. Verification of annotated data was done by the postgraduate student in the NLP domain. The whole process was approved by their professor, who is a domain expert. The whole process is shown in Figure 4.

Annotators Training: Annotators had a training session on annotation, and the guidelines mentioned above were provided to them.

Inter-Annotator Agreement: Before the actual annotation process started, an inter-annotator agreement analysis was conducted over the same data from 200 reviews. This agreement gives a measure of the extent to which the annotation process is reproducible or consistent [18]. Also, this agreement ensures that the annotators have a common and complete understanding of the requirements and indicates the areas of disagreement that might require the annotators to provide more information [19]. To access the agreement on categorical data, the Kappa statistic is used. There are two variations in Kappa's statistics, "Cohen's kappa and Fleiss' kappa." Both are almost similar but only differ in one important aspect. Fleiss 'kappa allows the metrics to be calculated for several annotators [20]. While Cohen's kappa limits the number of annotators to two [21]. As our annotation process was done by two annotators, we used Cohen's Kappa statistics for our work, and the result is shown in Table 2 and Table 3 respectively.

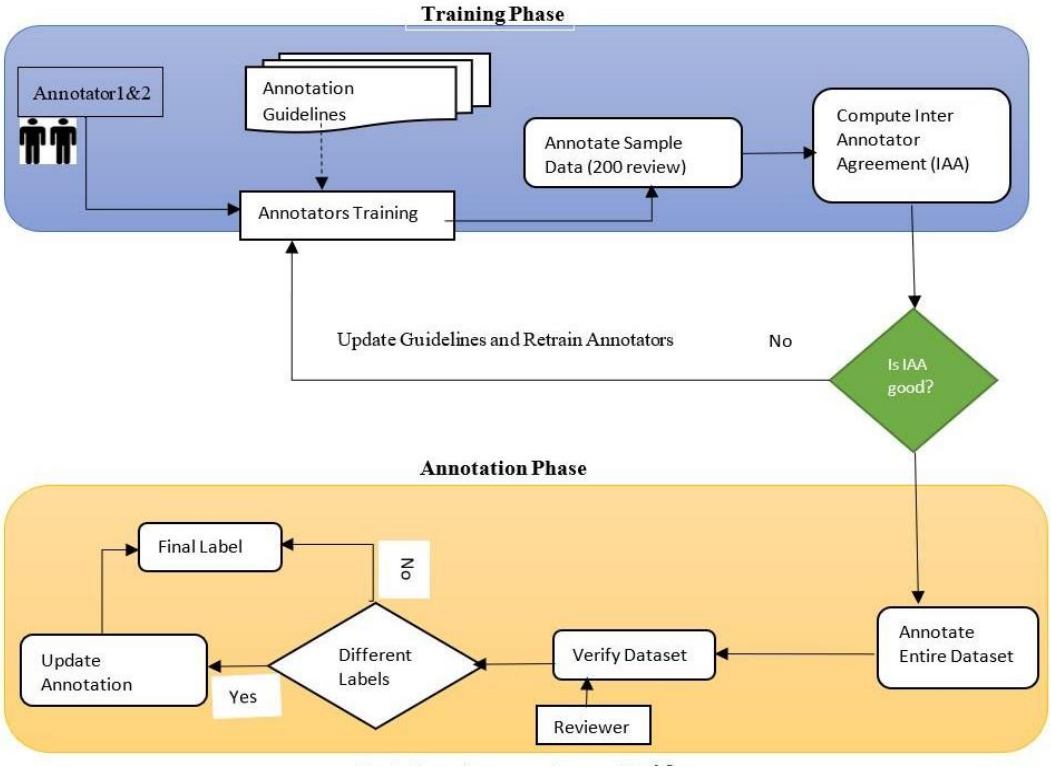

Figure 4: Data Construction Pipeline for Home Service App Reviews

Table 2:Inter-Annotator Agreement between annotators for Opinion Extraction

Opinion Term Extraction (OTE) agreement	91.92%
Aspect Category Detection (ACD) agreement:	84.65%
Aspect Sentiment Classification ASC agreement	94.38%

Table 3:Category-Wise Correlation among annotators

Category	TT	FF	TF	FT
company service	291	194	26	13
app_utility	100	412	2	10
service_schedule	8	515	1	0
service_quality	22	480	13	9
app_ui	10	511	3	0
customer_support	5	516	1	2
service_person_attitude	27	481	2	14
cost	10	514	0	0
App functionality	3	521	0	0

IV. RESULTS AND DISCUSSION

This study examined all three subtasks. We started with the pre-processing of the dataset. After this, we extracted features from the data and finally performed all ABSA tasks individually using some popular machine learning models. Moreover, [22] conducted detailed experiments on this benchmark dataset. They proposed a novel pattern-based approach for extracting multiple opinion terms. In addition, the identified aspects were organized into predefined general categories and subsequently classified according to their polarity using both machine learning and deep learning techniques.

A. Pre-processing & Feature Engineering

In this study, standard pre-processing techniques were applied to the dataset. These steps included the removal of stop words and punctuation, conversion of all text to lowercase, and stemming. After pre-processing, tokenization was performed to extract individual words from each review, which were then used to construct a vocabulary. Based on this vocabulary, a feature matrix was generated in which each review was represented as a vector. To compute the feature values for each term in the corpus, the Term Frequency–Inverse Document Frequency (TF-IDF) method was employed.

B. Results

The extracted features were used to train several widely adopted supervised machine learning algorithms, including Support Vector Machines (SVM), k-Nearest Neighbours (KNN), and Naïve Bayes. The dataset was divided into training and testing sets, with 70% of the data used for training and the remaining 30% reserved for testing. All three subtasks of ABSA were tested. **Table *4*** shows the comparisons of the machine learning models for opinion terms extraction. The opinion terms extraction task was treated as a single task where the original sentence and the output were extracted terms. Results show that SVM achieved the highest accuracy of 74.91% than others; however, its low precision 36.99% and recall indicate its limited effectiveness in correctly identifying sentences containing the target term. Therefore, it seems the model is biased towards the majority class and fails to capture true positive instances.

Table 4: Result Analysis for Opinion Terms Extraction

Methods	Accuracy	Precision	Recall	F1-Score
SVM	0.74	0.36	0.25	0.28
KNN	0.70	0.30	0.28	0.29
Naive Bayes	0.65	0.15	0.09	0.09

A more balanced performance is demonstrated by the KNN model, achieving the highest recall 28.24% and F1-score 29.15% among all evaluated methods. Its accuracy 70.11% (slightly lower than that of SVM), and improved recall indicates a better capability to detect sentences in which the target opinion term is present, making it more suitable for this task. For the OTE, the Naive Bayes classifier exhibited the weakest performance across all evaluation metrics, with a lowest precision 15.43% and recall 9.04%, suggesting that the strong independence assumptions of Naive Bayes confined its effectiveness when applied to sparse and short text data such as individual sentences.

Table ***5*** discusses the results of Aspect categorization, which is the second subtask of ABSA. Aspect categorization was a composite task where the input sentence and the extracted opinion term were considered as input, and the aspect category was predicted. As the category was predicted using a composite input, and it is a multiclass problem, overall, this task was relatively challenging, and all the evaluation matrices produced relatively low scores. In terms of accuracy, again, the SVM dominated other models; however, low precision and recall show that it fails to correctly assign aspect categories to a large proportion of sentences. Recall of KNN is comparable, but again, its low precision indicates that it struggles to clearly separate predefined aspect categories with composite input. Naïve Bayes performed worse than all for multiclass classification problems.

Results of the third subtask, Aspect Sentiment Classification, are presented in Table 6. Sentiment classification is a binary classification problem. SVM and KNN, both classifiers, achieved an accuracy of 91%, outperforming Naive Bayes, which obtained an accuracy of 88%. The discrepancy in precision, Recall and F1 score suggests that the dataset is likely imbalanced, and the classifiers tend to favour the majority polarity class, leading to inflated accuracy values.

Table 5: Result Analysis for the Aspect of Categorization

Methods	Accuracy	Precision	Recall	F1-Score
SVM	0.91	0.495	0.495	0.495
KNN	0.91	0.467	0.467	0.467
Naive Bayse	0.88	0.486	0.4867	0.4867

Table 6: Result Analysis for Aspect Sentiment Classification

Methods	Accuracy	Precision	Recall	F1-Score
SVM	0.589	0.22	0.23	0.22
KNN	0.549	0.18	0.22	0.19
Naive Bayes	0.334	0.035	0.039	0.034

V. CONCLUSION

This paper presents a novel domain-specific dataset (HS-ABSA) for aspect-based sentiment analysis for the service industry. Since most publicly available ABSA datasets are product-oriented, such as laptops and cameras, this dataset was specifically curated to capture the distinct features of service-related user feedback. For this purpose, reviews were collected from the mobile application of a real-world service provider (URBAN CLAP) offering home maintenance and related services. Python libraries were used to scrape reviews. A total of 4,164 reviews were scraped and annotated based on the guidelines provided by SemEval Task 4 (2014) and Android mobile application guidelines.

HS-ABSA has been prepared to cover different research tasks of ABSA. Three tasks have been considered, including opinion term extraction, aspect category detection, and sentiment classification. Moreover, HS-ABSA provides baseline results to evaluate the mentioned tasks. In the future, we aim to extend HS-ABSA to other service reviews and add additional domains. We also aim to increase the dataset size so that baseline model results can be improved.

REFERENCES

[1] S. Elmitwalli and J. Mehegan, "Sentiment analysis of COP9-related tweets: a comparative study of pre-trained models and traditional techniques," *Front. Big Data*, vol. 7, p. 1357926, Mar. 2024, doi: 10.3389/fdata.2024.1357926.

[2] E. Cambria, B. Schuller, Y. Xia, and C. Havasi, "New Avenues in Opinion Mining and Sentiment Analysis," *IEEE Intell. Syst.*, vol. 28, no. 2, pp. 15–21, Mar. 2013, doi: 10.1109/MIS.2013.30.

[3] B. Liu and L. Zhang, "A Survey of Opinion Mining and Sentiment Analysis," in *Mining Text Data*, C. C. Aggarwal and C. Zhai, Eds., Boston, MA: Springer US, 2012, pp. 415–463. doi: 10.1007/978-1-4614-3223-4_13.

[4] Q. Liu, Z. Gao, B. Liu, and Y. Zhang, "A Logic Programming Approach to Aspect Extraction in Opinion Mining," in *2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*, Atlanta, GA, USA: IEEE, Nov. 2013, pp. 276–283. doi: 10.1109/WI-IAT.2013.40.

[5] F. Xianghua, L. Guo, G. Yanyan, and W. Zhiqiang, "Multi-aspect sentiment analysis for Chinese online social reviews based on topic modeling and HowNet lexicon," *Knowl.-Based Syst.*, vol. 37, pp. 186–195, Jan. 2013, doi: 10.1016/j.knosys.2012.08.003.

[6] M. Hu and B. Liu, "Mining and summarizing customer reviews," in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, Seattle WA USA: ACM, Aug. 2004, pp. 168–177. doi: 10.1145/1014052.1014073.

[7] Tun Thura Thet, J.-C. Na, and C. S. G. Khoo, "Aspect-based sentiment analysis of movie reviews on discussion boards," *J. Inf. Sci.*, vol. 36, no. 6, pp. 823–848, Dec. 2010, doi: 10.1177/0165551510388123.

[8] N. Alturaief, H. Aljamaan, and M. Baslyman, "AWARE: Aspect-Based Sentiment Analysis Dataset of Apps Reviews for Requirements Elicitation," in *2021 36th IEEE/ACM International Conference on Automated Software Engineering Workshops (ASEW)*, Melbourne, Australia: IEEE, Nov. 2021, pp. 211–218. doi: 10.1109/ASEW52652.2021.00049.

[9] M. Pontiki, D. Galanis, J. Pavlopoulos, H. Papageorgiou, I. Androutsopoulos, and S. Manandhar, "SemEval-2014 Task 4: Aspect Based Sentiment Analysis," in *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, Dublin, Ireland: Association for Computational Linguistics, 2014, pp. 27–35. doi: 10.3115/v1/S14-2004.

[10] M. Pontiki *et al.*, "SemEval-2016 Task 5: Aspect Based Sentiment Analysis," in *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, San Diego, California: Association for Computational Linguistics, 2016, pp. 19–30. doi: 10.18653/v1/S16-1002.

[11] M. Al-Smadi, O. Qawasmeh, M. Al-Ayyoub, Y. Jararweh, and B. Gupta, "Human Annotated Arabic Dataset of Book Reviews for Aspect Based Sentiment Analysis," *Int. J. Comput. Linguist. Appl.*, vol. 6, no. 2, pp. 1–16, 2015.

[12] O. Alqaryouti, N. Siyam, and K. Shaalan, "A Sentiment Analysis Lexical Resource and Dataset for Government Smart Apps Domain," in *Proceedings of the International Conference on Advanced Intelligent Systems and Informatics 2018*, vol. 845, A. E. Hassanien, M. F. Tolba, K. Shaalan, and A. T. Azar, Eds., in Advances in Intelligent Systems and Computing, vol. 845. , Cham: Springer International Publishing, 2019, pp. 230–240. doi: 10.1007/978-3-319-99010-1_21.

[13] S. Rani and W. Anwar, "Resource Creation and Evaluation of Aspect Based Sentiment Analysis in Urdu," *J. King Saud Univ. – Comput. Inf. Sci.*, vol. 32, no. 9, pp. 1031–1042, 2020.

[14] G. Kontonatsios, N. Spanos, D. Chatzakou, and H. Karanikas, "FABSA: An Aspect-Based Sentiment Analysis Dataset of User Reviews," *Data Brief*, vol. 48, p. 109143, 2023.

[15] M. Pontiki, D. Galanis, H. Papageorgiou, I. Androutsopoulos, S. Manandhar, and M. Al-Smadi, "SemEval-2016 Task 5: Aspect Based Sentiment Analysis," presented at the Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016), San Diego, California: Association for Computational Linguistics, 2016, pp. 19–30.

[16] M. Saeidi, G. Bouchard, M. Liakata, and E. Shutova, "SentiHood: Targeted Aspect Based Sentiment Analysis Dataset for Urban Neighbourhoods," presented at the Proceedings of COLING 2016: The 26th International Conference on Computational Linguistics, Osaka, Japan, 2016, pp. 1546–1556.

[17] Q. Jiang, L. Chen, R. Xu, X. Ao, and M. Yang, "A Challenge Dataset and Effective Models for Aspect-Based Sentiment Analysis," presented at the Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP), Hong Kong, China: Association for Computational Linguistics, 2019, pp. 6280–6285.

[18] N. Alturaief, M. Al-Razgan, H. Al-Khalifa, and L. Al-Sharif, "AWARE: Aspect-Based Sentiment Analysis Dataset of App Reviews for Requirements Elicitation," in *Proceedings of the 29th IEEE International Requirements Engineering Conference (RE)*, Notre Dame, Indiana, USA: IEEE, 2021, pp. 149–159.

[19] R. Artstein, "Inter-annotator Agreement," in *Handbook of Linguistic Annotation*, N. Ide and J. Pustejovsky, Eds., Dordrecht: Springer Netherlands, 2017, pp. 297–313. doi: 10.1007/978-94-024-0881-2_11.

[20] P. Takala, P. Malo, A. Sinha, and O. Ahlgren, "Gold-standard for Topic-specific Sentiment Analysis of Economic Texts," in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, and S. Piperidis, Eds., Reykjavik, Iceland: European Language Resources Association (ELRA), May 2014, pp. 2152–2157. [Online]. Available: http://www.lrec-conf.org/proceedings/lrec2014/pdf/1021_Paper.pdf

[21] J. L. Fleiss, B. Levin, and M. C. Paik, *Statistical methods for rates and proportions*, 3rd ed. in Wiley series in probability and statistics. Hoboken, N.J: J. Wiley, 2003.

[22] J. Cohen, "A Coefficient of Agreement for Nominal Scales," *Educ. Psychol. Meas.*, vol. 20, no. 1, pp. 37–46, Apr. 1960, doi: 10.1177/001316446002000104.

[23] A. Maroof, S. Wasi, S. I. Jami, and M. S. Siddiqui, "Aspect-Based Sentiment Analysis for Service Industry," *IEEE Access*, vol. 12, pp. 109702–109713, 2024, doi: 10.1109/ACCESS.2024.3440357.