Aspect-based Sentiment Analysis for Vietnamese Reviews about Fashion Product on E-commerce Websites

DUONG VAN HIEP, FPT University, Vietnam
PHAM XUAN KHANG, FPT University, Vietnam
HOANG THAI DUONG, FPT University, Vietnam
TRAN TUAN ANH, FPT University, Vietnam
MAI ANH TRUONG, FPT University, Vietnam

Millions of reviews are generated on e-commerce platforms, and deeply analyzing them can provide valuable insights for both sellers and buyers. This paper addresses the aspect-based sentiment analysis problem to evaluate the aspects and sentiments of Vietnamese reviews related to fashion products on e-commerce websites. The contributions of this paper are three-fold. Firstly, we introduce a dataset comprising 4,539 reviews on fashion products, with a total of 6,580 aspect-sentiment pairs. Additionally, we conduct baseline experiments applying aspect-based sentiment analysis methods, utilizing machine learning for single-task learning and deep learning for multitask learning. Hyperparameter optimization techniques are employed to improve model performance. For aspect detection, the Support Vector Machine (SVM) model achieved an F1-score of 0.86, while for sentiment classification, the BiGRU+Conv1D model obtained an F1-score of 0.91. These results highlight the effectiveness of the proposed models and demonstrate the advantages of utilizing multitask deep learning architectures for sentiment classification. Furthermore, a comprehensive comparison between single-task and multitask learning approaches is conducted to determine the most suitable method for predicting aspects and sentiments in fashion reviews. Experimental results indicate that the BiGRU+Conv1D model in the multitask learning approach outperforms other models, achieving notable accuracy in both aspect detection and sentiment classification tasks.

CCS Concepts: • Computing methodologies → Natural language processing; Machine learning; Deep learning; • Information systems → Fashion data.

Additional Key Words and Phrases: Natural Language Processing, Machine Learning, Deep Learning, Fashion Data, Aspect-Based Sentiment Analysis

ACM Reference Format:
DUONG VAN HIEP, PHAM XUAN KHANG, HOANG THAI DUONG, TRAN TUAN ANH, and MAI ANH TRUONG. 2018. Aspect-based Sentiment Analysis for Vietnamese Reviews about Fashion Product on E-commerce Websites. Proc. ACM Meas. Anal. Comput. Syst. 37, 4, Article 111 (August 2018), 12 pages. https://doi.org/XXXXXXX.XXXXXXX

Authors' addresses: DUONG VAN HIEP, hiepdvhe181185@fpt.edu.vn, FPT University, FPT University, Hanoi, Thach That, Vietnam, 1000; PHAM XUAN KHANG, FPT University, FPT University, Hanoi, Vietnam, khangpxhe180735@fpt.edu.vn; HOANG THAI DUONG, FPT University, FPT University, Hanoi, Vietnam, duonghthe181802@fpt.edu.vn; TRAN TUAN ANH, FPT University, FPT University, Hanoi, Vietnam, anhtthe181406@fpt.edu.vn; MAI ANH TRUONG, FPT University, FPT University, Hanoi, Vietnam, truongmahe182474@fpt.edu.vn.

2476-1249/2018/8-ART111 $15.00
https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Aspect-based sentiment analysis (ABSA) is an important task in Natural Language Processing (NLP). Instead of merely extracting the overall sentiment polarity of the text (positive, negative, neutral) in sentiment analysis, ABSA delves deeper by displaying the polarity of each aspect of the text. In review comments on e-commerce platforms, a single review often contains multiple aspects, making ABSA more powerful for analyzing sentiment polarity than traditional sentiment analysis. For example, in product reviews for fashion items like clothing, customers do not always provide an overall assessment of the product. Instead, they tend to evaluate specific aspects of the item, such as price, appearance, fabric quality, etc. A customer might like the appearance but feel frustrated if the fabric quality is poor or uncomfortable with the store's customer service. This is why ABSA is truly essential for gaining deeper insights into what customers think about the products they buy. With the explosion of e-commerce, millions of reviews are generated daily by customers. This valuable source of information not only helps sellers understand what customers like and dislike about their products, allowing them to address limitations and enhance competitiveness, but it also serves other customers who can read and consider previous buyers' experiences before making a purchase.Among the vast array of products available on e-commerce platforms, fashion items are always a top priority for customers due to their practicality and usefulness. Additionally, fashion product reviews, especially for clothing, encompass various aspects, from price and fabric quality to appearance, enabling deeper analyses of customer attitudes toward clothing. This is why we chose clothing reviews to build a dataset aimed at addressing the ABSA challenge within e-commerce reviews.

This paper makes three main contributions. First, we collected a new Vietnamese dataset of lipstick reviews on e-commerce platforms for the aspect-based sentiment analysis task. To our knowledge, there is no specific dataset for Vietnamese fashion product reviews.Second, we propose an architecture that combines two effective machine learning and deep learning models to detect aspects and their corresponding sentiments in reviews. Finally, a comprehensive comparison between single-task learning and multi-task learning is conducted to determine the best approach for aspect-based sentiment analysis.

2 RELATED WORKS

With a growing body of work and many workshops targeted at improving methods and approaches, aspect-based sentiment analysis (ABSA) remains a major field of study. SemEval 2014 (Pontiki et al., 2014), SemEval 2015 (Pontiki et al., 2015), SemEval 2016 (Pontiki et al., 2016), VLSP 2018 (Nguyen et al., 2018), and, more recently, SemEval 2020 (Schneider et al., 2020) and SemEval 2021 (Zhang et al., 2021) are important events. These workshops have covered a range of ABSA topics, including sentiment classification and aspect category identification.

Both single-task and compound ABSA techniques are recent developments. Subtasks like sentiment categorization and aspect category identification are part of single-task ABSA. While Wang et al. (2016) used attention-based LSTM for aspect sentiment classification, Zhou et al. (2015) used representation learning to get an F1-score of 90.10 for aspect category detection on the SemEval 2014 dataset using representation learning, while Wang et al. (2016) utilized attention-based LSTM for aspect sentiment classification. More recent work has explored transformer-based models, such as BERT and its derivatives, which have significantly improved performance across various ABSA tasks.

Researchers have also made advancements in compound ABSA, creating models that extract aspects and their moods at the same time. While more recent models use multi-task learning frameworks to improve the extraction process, He et al. (2019) suggested an interactive multi-task learning network (e.g., Liu et al., 2022).

Bảng 1. Aspect Definition

Aspect	Definition
PRICE	Reviews mention the cost of the clothes.
APPEARANCE	Reviews mention the presence of the clothes such as beautiful or ugly
DELIVERY	Reviews mention the delivery service such as the time, the shipper's attitude
FABRIC QUALITY	Reviews mention the fabric of the clothes such as good or not.
PACKAGING	Reviews mention the quality of packing, whether the clothes are well packed or not.
SERVICE	Reviews mention the service of the clothes.
SIZE FIT	Reviews mention the flexible of the clothes are well fit or not.
OTHERS	Spam reviews

In the Vietnamese context, the availability of datasets for sentiment analysis has expanded. The SA-VLSP2018 dataset, which focuses on hotels and restaurants, was introduced by Nguyen et al. (2018). More recent datasets include UIT-ViSD4SA, which Nguyen et al. (2021) used to achieve an F1-score of 62.76% through span detection techniques. Additionally, Nguyen et al. (2018) created the UIT-VSFC dataset, containing over 16,000 sentences of student feedback.

3 DATA SET

3.1 Task Definition

We built a Vietnamese dataset for the aspect-based sentiment analysis task. This dataset contains 4,539 Vietnamese reviews about clothes from Tiki. There are 2 sub-tasks in this dataset: aspect detection and sentiment classification. In the aspect detection sub-task, we focus on finding the aspects mentioned in the review. There are 7 aspects: PRICE, APPEARANCE, DELIVERY, FABRIC QUALITY, PACKAGING, SERVICE, SIZE FIT and aspect indicating spam: OTHERS. Table 1 shows the definition of all aspects. Another sub-task is classifying the polarity of these aspects into: Positive, Neutral or Negative. We split the dataset into three sets: 3,631 reviews for training, 454 reviews for validation, and 454 reviews for testing. The training and validation will help us to build detection and classification models, and the test set is used to evaluate the performance of the models

3.2 Annotation process

We divided our annotation process into two main phases: the training phase and the labeling phase. The training phase consists of five sub-phases, each containing 200 reviews. To measure inter-annotator agreement and assess annotation quality, we use Cohen's Kappa score (Cohen, 1960). When the score between annotators in each sub-phase exceeds 0.65, we conclude the training for that sub-phase and proceed to the next.

Once the training phase is complete, annotators move on to label the remaining dataset in the labeling phase. This phase is divided into two sub-phases, each containing approximately 1,000 reviews. Three annotators independently label 1,000 reviews each. For reviews with disagreements, we automatically select the label chosen by 2/3 annotators.

Bảng 2. Several examples of the dataset.

Review	Aspect	Polarity
hàng đẹp chất vải mát, giao hàng nhanh nhưng đóng gói xấu (Beautiful product, quick arrival and bad packaging)	APPEARANCE DELIVERY PACKAGING	APPEARANCE: Positive DELIVERY: Positive PACKAGING: Negative
giao hàng chậm (low delivery)	DELIVERY	DELIVERY: Negative
Size rất vừa, giá rẻ, nhân viên nhiệt tình (very fit, affordable and kind staff)	SIZE FIT PRICE SERVICE	SIZE FIT: Positive Price: Positive SERVICE: Positive

	Review	Avg aspect/ review	Positive	Neutral	Negative	Total sentiment
Train	3631	1.458276	3584	940	771	5295
Val	454	1.389868	412	119	100	631
Test	454	1.440529	451	124	79	654

Fig. 1. Statistics about the experimental dataset.

3.3 Statistics

Our dataset contains 4,539 reviews, including 7 sentiment aspects and 1 aspect indicating spam or unrelated reviews.

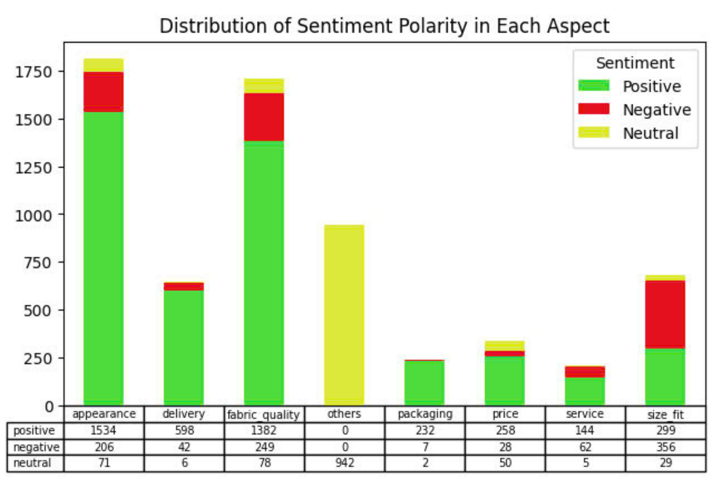

Fig. 2. Distribution of sentiment polarity in each aspect

Overall, positive sentiment accounts for a very large proportion in all aspects of our dataset, except for the OTHERS and SIZE FIT aspects. Additionally, the majority of reviews in the dataset focus on APPEARANCE and FABRIC QUALITY, comprising over one-third of the total reviews. This suggests that consumers prioritize the appearance and fabric quality of clothing, as these factors directly impact their user experience. The greatest imbalance between sentiments is most apparent in three aspects: DELIVERY, PACKAGING, and SERVICE. This can be attributed to the recent rapid development of shipping services, which has significantly reduced delivery times. Meanwhile, PACKAGING and SERVICE are elements that sellers can control and manage to create a positive first impression for buyers.

4 MACHINE LEARNING APPROACH

This section gives information about the model that is used for the aspect-based sentiment analysis problem. The model has three main components: input layer & hidden layers, output layer for single-task learning, and output layer for multi-task learning approach. For the input layer and hidden layers in subsection 4.1, the model architecture is the same between single task learning and multi-task learning. Based on the learning approach, the output layer can be different. The subsection 4.2 and 4.3 give more details about the output layers of two approaches.

4.1 Aspect detection model

In this section, we will start by embedding the training data set in text format (processed_text) using TF-IDF. Our output matrix has dimensions (3630, 716), where 3630 is the number of comments in the training set, and 716 is the number of tokens detected by the model.

TF:

$$\mathrm{TF}(t, d) = \frac{f(t, d)}{\max\{f(w, d) : w \in d\}}$$

IDF:

$$\mathrm{IDF}(t, D) = \log\left(\frac{|D|}{|\{d \in D : t \in d\}|}\right)$$

The reason we chose to use this method is that it is simple, and our dataset is not too large. Next, we chose to use Random Forest to classify the aspects mentioned in each comment.

Then, we use Random Forest to train a model to classify these vectors into 7 aspects: PRICE, APPEARANCE, DELIVERY, FABRIC QUALITY, PACKAGING, SERVICE, SIZE FIT. We split the data into training and testing sets to ensure accuracy during evaluation. After training, we continue to use the trained model to make predictions and check the prediction results with the testing data.

Our results show that The model performs very well for aspects like "packaging,""delivery,"and "service,"but struggles to detect aspects such as "price"and "service"(with a low F1-score due to low recall).

We choose to use Random Forest for our aspect detection dataset because this model has strong and flexible classification capabilities, handles high-dimensional data from TF-IDF well, and minimizes overfitting by combining multiple decision trees. Random Forest is also effective at reducing noise, easy to tune for optimal accuracy, and highly reliable, enabling the model to accurately identify aspects like PRICE, APPEARANCE, DELIVERY, and other aspects in your data.

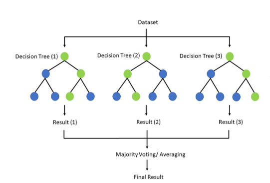

Fig. 3. Random forest model

4.2 Sentiment Analysis Model

To address our problem, after using TF-IDF and Random Forest for aspect detection, we implemented an SVM (Support Vector Machine) for the sentiment classification model, categorizing sentiment into positive, neutral, and negative classes.

SVM was chosen for its strengths in handling high-dimensional data, its robustness with smaller datasets, and its ability to create a clear margin of separation between different classes, reducing misclassification. It is particularly well-suited for text data transformed by TF-IDF, where the feature space can be large and complex.

The SVM model aims to find a hyperplane that best separates the classes in the training data. The key optimization problem in SVM is to minimize the following:

$$\min \frac{1}{2}\|w\|^2$$

subject to:

$$y_i(w \cdot x_i + b) \geq 1, \quad \forall i = 1, 2, \ldots, n$$

where:

- w represents the weight vector.
- x_i represents the input feature vector for the i-th training example.
- y_i represents the class label ($+1$ or -1) for the i-th training example.
- b is the bias term.

For non-linearly separable data, we introduce slack variables ξ_i to allow some misclassifications, resulting in a soft-margin SVM:

$$\min \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i$$

subject to:

$$y_i(w \cdot x_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0$$

where:

- C is the regularization parameter, controlling the trade-off between maximizing the margin and minimizing classification errors.
- ξ_i represents slack variables that allow for some misclassification in the training data.

SVM uses different kernel functions to map the input space into a higher-dimensional feature space, allowing the model to handle non-linear relationships between features. In our implementation, both linear and Radial Basis Function (RBF) kernels were explored.

4.3 Hyperparameter Optimization

In the third step, we perform hyperparameter optimization using GridSearchCV to determine the best configuration for the SVM model. This step ensures that the model is tuned for optimal performance.

We defined a grid of hyperparameters to search over, including:

- C: Regularization parameter.
- γ: Kernel coefficient for the RBF kernel.
- kernel: Type of kernel function (e.g., 'linear', 'rbf').

GridSearchCV performs cross-validation (in our case, cv = 5) to evaluate the model's performance for each combination of hyperparameters. The metric used for scoring was the weighted F1-score:

$$\text{F1-score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

This metric helps in evaluating the model's ability to correctly classify both major and minor classes, particularly in imbalanced datasets.

For each aspect, we trained an SVM model using GridSearchCV and selected the model with the best parameters. These parameters were used to evaluate the model on the validation set, and the results showed that the model worked well for some aspects but struggled with others due to lower class diversity in the training data.

Through hyperparameter optimization, we refined our model, ensuring that it performs well on unseen data across different aspects in the dataset. The process of fine-tuning parameters such as C, γ, and the kernel function helped to strike the right balance between model complexity and prediction accuracy.

5 DEEP LEARNING APPROACH

Our deep learning approach utilizes the most effective model architecture for multitasking Aspect-Based Sentiment Analysis (ABSA), as discussed in (Tran et al., 2023). This architecture is designed to be simplified, making it well-suited for small datasets while still achieving strong performance across classification tasks. An illustration of the architecture is provided in Figure 4.

5.1 Input and Hidden layers

First, in input layers we use embedding to help the model learn the semantic features of words in a sentence rather than just treating words as a sequence of characters. This allows the model to understand words with similar meanings, thereby improving its classification and aspect recognition capabilities.

Then, in hidden layers we use BiGRU gathers context from both before and after a word. This helps the model better capture the semantic relationships related to each aspect.For example, the word "tốt"in the context of 'cheap price' belong to PRICE aspect, while in the context of "fabric feel when using", it falls under the FABRIC QUALITY aspect. Thanks to BiGRU, the model can make more accurate distinctions based on the full context.

The GRU for each time step is calculated as follows:

$$z_t = \sigma(W_z \cdot h_{t-1} + U_z \cdot x_t)$$

$$r_t = \sigma(W_r \cdot h_{t-1} + U_r \cdot x_t)$$

$$\tilde{h}_t = \tanh(W \cdot (r_t * h_{t-1}) + U \cdot x_t)$$

$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t$$

where z_t is the update gate, r_t is the reset gate, and h_t is the hidden state.

For BiGRU, we use forward and backward GRUs:

$$\overrightarrow{h}_t = \text{GRU}(x_t), \quad \overleftarrow{h}_t = \text{GRU}(x_t)$$

$$h_t = [\overrightarrow{h}_t; \overleftarrow{h}_t]$$

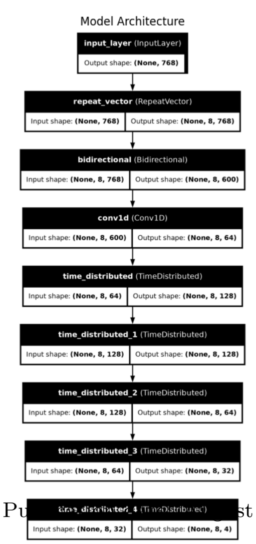

Fig. 4. BiGRU model architecture

After the BiGRU layer, the model has captured the complete context and hidden aspects within some complex comments.

Finally, the Conv1D layer helps the model identify semantically significant phrases by applying filters (kernels) to the output sequence from the BiGRU layer. Each filter detects local patterns (n-grams) within the embedding sequences, such as phrases or groups of words relevant to each aspect. For example, phrases like "giá tốt", "chất vải"and "giao hàng nhanh"typically characterize specific aspects (like PRICE, FABRIC QUALITY, DELIVERY). Conv1D detects these patterns by scanning through the embedding sequence and recognizing prominent phrases, aiding in more accurate differentiation between aspects.

The formula for Conv1D is:

$$c_i = f(w \cdot h_{i:i+k-1} + b)$$

where:

- . w is the filter.
- . $h_{i:i+k-1}$ are the vectors from position i to $i + k - 1$ in the BiGRU sequence.
- . f is the activation function (e.g., ReLU).
- . b is the bias.

5.2 Output layer and Hyperparameter optimization

The Aspect-Specific Output Layer is designed to classify sentiments for eight product review aspects across four sentiment categories: None, Negative, Neutral, and Positive. The model applies a sequence of Bidirectional GRU and Conv1D layers to extract detailed, aspect-specific features. It then utilizes a TimeDistributed Dense layer with four units per aspect, activated by softmax, yielding an output shape of $(8, 4)$. Each aspect's sentiment is independently classified, allowing the model to assign unique labels for each aspect. For the classification task, categorical cross-entropy is used as the loss function to optimize the model.

Categorical cross-entropy, a commonly used loss function for multi-class classification, is mathematically defined as:

$$\text{Loss}_{\text{CCE}} = -\sum_{i=1}^{C} y_i \log(\hat{y}_i)$$

where:

- C is the total number of classes (4 in this model).
- y_i is the true label, represented as a one-hot encoded vector for each aspect.
- $\hat{y}_i$ is the predicted probability for each class, generated by the softmax function.

This loss penalizes incorrect predictions more heavily, as the logarithmic function magnifies errors when $\hat{y}_i$ diverges from y_i, encouraging the model to improve classification accuracy for each sentiment aspect.

Hyperparameter optimization is performed using Keras Tuner's RandomSearch, specifically tuning the model's GRU units, learning rate, and dropout rate to maximize performance. The tuner iterates over possible values, selecting the optimal configurations of 300 GRU units, a learning rate of 0.0001, and a dropout rate of 0.3, based on the lowest validation loss. This approach yielded an average sentiment F1-score of 0.91, reflecting significant improvements in classification accuracy.

6 RESULT AND EVALUATION

6.1 Evaluation metrics

Due to the imbalance of aspect and sentiment categories in the experimental dataset, the F1-score is used to evaluate model performance. This metric accounts for the imbalance, providing a more accurate view of the predictive models' effectiveness. We denote $F1_{ad}$ and $F1_{sc}$ as the F1-scores for the aspect detection and sentiment classification tasks, respectively. The formula for the F1-score, as presented by Sokolova and Lapalme (2009), is shown below, where n is the number of samples, and TP, FP, and FN represent True Positives, False Positives, and False Negatives, respectively:

$$\text{Precision} = \frac{\sum_{i=1}^{n} \text{TP}_i}{\sum_{i=1}^{n}(\text{TP}_i + \text{FP}_i)}$$

$$\text{Recall} = \frac{\sum_{i=1}^{n} \text{TP}_i}{\sum_{i=1}^{n}(\text{TP}_i + \text{FN}_i)}$$

$$\text{F1-score} = \frac{2 \times (\text{Precision} \times \text{Recall})}{\text{Precision} + \text{Recall}}$$

6.2 Results and Discussion

6.2.1 Machine Learning Model (Random Forest for Aspect Detection and SVM for Sentiment Analysis).

In Aspect Detection, the Random Forest model achieved an overall macro F1-score of 0.86 for aspect detection. Individual aspect F1-scores were notably high for categories such as delivery (0.96) and packaging (0.97), indicating that Random Forest is effective in identifying aspects with clearly distinguishable features. However, aspects such as service (0.71) and size_fit (0.80) displayed lower performance, likely due to class imbalance or less distinct feature separability.

	Aspect Detection (F1 Score)	Sentiment Analysis (F1 Score)
appearance	0.90	0.77
delivery	0.96	0.95
fabric_quality	0.88	0.66
others	0.84	NaN
packaging	0.97	0.92
price	0.85	0.76
service	0.71	0.76
size_fit	0.80	0.55
overall	0.86	0.76

Fig. 5. F1 score of each model machine learning approach

In Sentiment Analysis, the SVM classifier, optimized for sentiment detection, attained an overall weighted F1-score of 0.76 across all aspects. High performance was observed for delivery (0.95) and packaging (0.92), while aspects such as size_fit (0.55) and fabric_quality (0.66) exhibited lower scores, potentially due to limited data or minimal variation in sentiment expressions for these aspects.

6.2.2 Deep Learning Model (GRU with Conv1D).

In Aspect Detection, the Bidirectional GRU with Conv1D model achieved an overall macro F1-score of 0.72. While this performance was slightly lower than that of the Random Forest model, the GRU model showed strong F1-scores for aspects such as delivery (0.92) and fabric_quality (0.82). However, aspects with more ambiguous sentiment expressions, such as service (0.33) and size_fit (0.62), had lower detection accuracy, indicating challenges in extracting nuanced aspect features.

In Sentiment Analysis, for sentiment classification, the GRU model achieved an overall weighted F1-score of 0.91, surpassing the SVM model. Particularly high F1-scores were obtained for packaging (0.98), service (0.96), and delivery (0.97), demonstrating the model's ability to capture complex sentiment nuances within these aspects. The machine learning approach yielded strong results in aspect detection, particularly for aspects with well-defined features. However, the deep learning model outperformed in sentiment classification, achieving a higher overall F1-score. This suggests that deep learning architectures may be more effective in capturing nuanced sentiment expressions, whereas traditional models such as Random Forest are highly effective in detecting clearly distinguishable aspects.

	Aspect	Aspect Detection (F1 Score)	Sentiment Analysis (F1 Score)
0	appearance	0.80	0.82
1	delivery	0.92	0.97
2	fabric_quality	0.82	0.82
3	others	0.74	0.89
4	packaging	0.69	0.98
5	price	0.57	0.94
6	service	0.33	0.96
7	size_fit	0.62	0.88
8	overall	0.72	0.91

Fig. 6. F1 score of each model deep learning approach

7 CONCLUSION AND FUTURE WORK

This paper addresses the aspect-based sentiment analysis of fashion product reviews on e-commerce platforms. In this paper, we introduce a new dataset containing 4,539 reviews about clothing on e-commerce platforms for the task of aspect-based sentiment analysis. For the task of predicting the aspects and sentiments of reviews, we compared two approaches: machine learning and deep learning. Specifically, the machine learning approach uses the TF-IDF model to transform the text, followed by Random Forest for aspect detection and SVM for sentiment classification. The deep learning approach, on the other hand, combines BiGRU and Conv1D layers to extract semantic features, with Keras Tuner used to optimize the model's hyperparameters.

The results show that the machine learning approach achieved more favorable results in terms of accuracy, though it requires greater complexity in implementation due to the combination of two models. This reflects a trade-off between accuracy and complexity. The highest F1-score was achieved by the BiGRU+Conv1D model in deep learning, with 91% for aspect detection and 72% for sentiment classification.

For future work, we are considering building an automated pipeline to collect reviews on e-commerce platforms for each product, process the reviews, predict their aspects and sentiments, and display the results on a dashboard. This pipeline will provide a broader view for sellers to better understand their products and for customers to make informed purchasing decisions.

8 REFERENCES

(1) Maria Pontiki, Dimitrios Galanis, John Pavlopoulos, Haris 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), Dublin, Ireland, Association for Computational Linguistics, pp. 27-35. https://doi.org/10.3115/v1/S14-2004.
(2) Maria Pontiki, Dimitrios Galanis, Haris Papageorgiou, Suresh Manandhar, Ion Androutsopoulos, John Pavlopoulos, and George Kontopoulos. 2015. SemEval-2015 Task 12: Aspect Based Sentiment Analysis. In Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), Denver, Colorado, Association for Computational Linguistics, pp. 486-495. https://doi.org/10.18653/v1/S15-2082.
(3) Maria Pontiki, Dimitrios Galanis, Haris Papageorgiou, Suresh Manandhar, and Ion Androutsopoulos. 2016. 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, pp. 19-30. https://doi.org/10.18653/v1/S16-1002.
(4) Kim Anh Nguyen, Xuan Bach Le, and Ngan Luu-Thuy Nguyen. 2018. Vietnamese Language and Speech Processing 2018: Named Entity Recognition and Aspect Based Sentiment Analysis. In Proceedings of the 5th International Workshop on Vietnamese Language and Speech Processing (VLSP 2018), Ho Chi Minh City, Vietnam
(5) Nathan Schneider, Manaal Faruqui, Matthew Lamm, Patrick Xia, Alexis Palmer, Graham Neubig, and Noah A. Smith. 2020. SemEval-2020 Task 5: Modelling Causal Reasoning in Language: Detecting, Grounding, and Categorizing Causes and Effects. In Proceedings of the 14th International Workshop on Semantic Evaluation (SemEval 2020), Barcelona, Spain, Association for Computational Linguistics, pp. 312-323. https://doi.org/10.18653/v1/2020.semeval-1.39.
(6) Xiang Zhang, Chenghua Lin, and Mark Stevenson. 2021. SemEval-2021 Task 4: Reading Comprehension of Abstract Meaning. In Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval 2021), Bangkok, Thailand, Association for Computational Linguistics, pp. 446-455. https://doi.org/10.18653/v1/2021.semeval-1.55.
(7) Siqiu Wang, Guangluan Xu, Jun Xu, and Hua Wu. 2016. Attention-Based LSTM for Aspect-Level Sentiment Classification. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP 2016), Austin, Texas, Association for Computational Linguistics, pp. 606-615. https://doi.org/10.18653/v1/D16-1058.
(8) Xiaowei Zhou, Xiaojun Quan, and Zhongang Qi. 2015. Multi-View Attention Networks for Aspect-Based Sentiment Analysis. In Proceedings of the 24th International Conference on Artificial Intelligence (IJCAI 2015), Buenos Aires, Argentina, pp. 3612-3618.

(9) Ruidan He, Wee Sun Lee, Hady W. Lauw, and Lidong Bing. 2019. An Interactive Multi-Task Learning Network for End-to-End Aspect-Based Sentiment Analysis. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019), Florence, Italy, Association for Computational Linguistics, pp. 504-515. https://doi.org/10.18653/v1/P19-1048.

(10) Qing Liu, Lin Gui, Yulan He, and Ruifeng Xu. 2022. BERT-Based Multi-Task Learning for Aspect-Based Sentiment Analysis. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL 2022), Dublin, Ireland, Association for Computational Linguistics, pp. 4195-4205. https://doi.org/10.18653/v1/2022.acl-long.298.

(11) Kim Anh Nguyen, Xuan Bach Le, and Ngan Luu-Thuy Nguyen. 2018. A Robust Approach to Vietnamese Named Entity Recognition and Aspect-Based Sentiment Analysis. In Proceedings of the 5th International Workshop on Vietnamese Language and Speech Processing (VLSP 2018), Ho Chi Minh City, Vietnam.

(12) J. Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20, 1, 37–46. DOI:https://doi.org/10.1177/001316446002000104.

(13) L. Breiman. 2001. Random forests. *Machine Learning*, 45, 1, 5–32. DOI:https://doi.org/10.1023/A:1010933404324

(14) C. Cortes and V. Vapnik. 1995. Support-vector networks. *Machine Learning*, 20, 3, 273–297. DOI:https://doi.org/10.1007/BF00994018

(15) J. Moody and C. J. Darken. 1989. Fast learning in networks of locally-tuned processing units. *Neural Computation*, 1, 2, 281–294. DOI:https://doi.org/10.1162/neco.1989.1.2.281

(16) Mark Alves. 2006. Linguistic Research on the Origins of the Vietnamese Language: An Overview. Journal of Vietnamese Studies 1, 2 (2006), 104–130.

(17) M. Sokolova and G. Lapalme. 2009. A systematic analysis of performance measures for classification tasks. *Information Processing Management*, 45, 4, 427–437. DOI:https://doi.org/10.1016/j.ipm.2009.03.002.

(18)

(19) Binyam Gebrekidan Gebre, Marcos Zampieri, Peter Wittenburg, and Tom Heskes. 2013. Improving Native Language Identification with TF-IDF Weighting. In Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications, pages 216–223, Atlanta, Georgia. Association for Computational Linguistics.

(20) Xueqing Liu and Chi Wang. 2021. An Empirical Study on Hyperparameter Optimization for Fine-Tuning Pre-trained Language Models. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 2286–2300, Online. Association for Computational Linguistics.

Received 29 October 2024; revised ; accepted