


Aspect-Based Sentiment Analysis on Amazon Product Reviews Using a Novel Hybrid Machine Learning Algorithm


Timothy Louis Scott

(Taylor's University, Subang Jaya, Selangor, Malaysia)

 <https://orcid.org/0009-0003-1561-7986>, timothylouisscott@sd.taylors.edu.my


Wei Wei Goh

(Taylor's University, Subang Jaya, Selangor, Malaysia)

 <https://orcid.org/0000-0002-8577-8481>, Weiwei.goh@taylors.edu.my

Navid Ali Khan

(Taylor's University, Subang Jaya, Selangor, Malaysia)

 <https://orcid.org/0009-0006-9305-2206>, Navidali.khan@taylors.edu.my

Abstract: On Amazon, buyers can submit reviews on products they have purchased. These reviews contribute to a potential buyer's decision-making process, as buyers read reviews to decide whether to buy a product. Additionally, sellers depend on reviews to improve their product offerings. Amazon's summary of reviews does not clearly indicate if an aspect of a product is mentioned positively or negatively. Buyers can manually read a small number of reviews to understand the overall sentiment towards a product, but reading reviews becomes progressively more difficult as the number of reviews increases, as it can lead to information overload. To address this problem, a hybrid machine learning classification algorithm that employs a branch of natural language processing, specifically aspect-based sentiment analysis, was developed to detect the polarity and key aspects mentioned in Amazon product reviews. Naïve Bayes, SVM, Decision Tree and Random Forest were compared to determine the two best algorithms for this purpose. The hybrid algorithm, named Soft Voting Hybrid Algorithm (SVHA), was implemented by training and testing a voting classifier using soft voting, which produced the final prediction by selecting the class with the highest average sum of probabilities from two base classifiers with the highest accuracies and macro F1-scores. Based on the experiments conducted, SVHA attained higher accuracies and macro F1-scores compared to the other four algorithms, showing its suitability in conducting aspect-based sentiment analysis.

Keywords: Sentiment analysis, E-commerce, Natural language processing, Machine learning, Product reviews

Categories: I.2.7

DOI: 10.3897/jucs.146032

1 Introduction

Electronic commerce (e-commerce) platforms allow users to buy and sell products online. E-commerce offers some advantages over physical stores. In the traditional commerce model, buyers need to travel to the store during its operating hours, whereas e-commerce allows buyers to shop at any time and from anywhere [Avlani (2020)].

Most e-commerce platforms allow buyers to leave reviews after purchasing a product. The number of reviews positively affects the number of sales for a product

[Park, Chung and Lee (2019)], as prospective buyers use existing reviews to make informed decisions [Nellutla et al. (2021)]. Product reviews are helpful to sellers as well, as sellers can identify the strengths and weaknesses of their products and can use buyers' feedback to improve their offerings [Zhang and Qiu (2021); Ramezani, Rahimi and Allan (2020)].

On many e-commerce platforms, it is difficult for sellers and buyers to easily and efficiently understand and analyse other buyers' sentiments on products being sold. Reading all the reviews is laborious [Jiang et al. (2021)], and it is difficult to identify positive, negative, or neutral aspects of a product in a review without reading it manually or using third-party tools. It is still feasible to read a small number of reviews, but not when the total number of reviews is high. This can lead to information overload, which negatively affects buyers' decision-making process [Hu and Krishen (2019)].

To address this problem, there is a need to develop a method to display a summary of product reviews that highlights important aspects of a product mentioned by reviewers, and their associated polarities. This study aimed to accomplish this through aspect-based sentiment analysis, where a hybrid algorithm named Soft Voting Hybrid Algorithm (SVHA) was developed based on a comparative study conducted between four existing algorithms (Naïve Bayes, SVM, Decision Tree and Random Forest). Rather than assigning a polarity to an entire review which may contain different aspects mentioned in a positive, negative or neutral manner, aspect-based sentiment analysis can assign a different polarity to each aspect in the review according to the sentiment expressed by the reviewer.

In relation to the above, the present study addressed three research questions.

1. How can aspect-based sentiment analysis be leveraged to improve the e-commerce shopping experience for buyers?
2. How can a hybrid aspect-based sentiment analysis algorithm be developed and implemented to efficiently process large volumes of product reviews from e-commerce platforms?
3. How does the proposed hybrid aspect-based sentiment analysis algorithm compare in effectiveness to existing approaches in processing and analysing product reviews?

This study aimed to accomplish the following three objectives, corresponding to the three research questions.

1. To conduct a comprehensive literature review of current research related to aspect-based sentiment analysis in e-commerce, identifying key gaps and challenges in existing studies.
2. To develop and implement a hybrid aspect-based sentiment analysis algorithm capable of processing large volumes of product reviews from e-commerce platforms.
3. To evaluate the effectiveness of the proposed hybrid aspect-based sentiment analysis algorithm against existing approaches.

This study contributes to the field of NLP and sentiment analysis by proposing a novel approach to aspect-based sentiment analysis, which can be harnessed in other domains beyond e-commerce. The development and implementation of the proposed approach can enhance the accuracy and efficiency of sentiment analysis, making it a valuable tool for businesses and researchers alike.

Moreover, this study presents an extensive literature review on sentiment analysis which can serve as a useful resource for researchers from academic and business environments, offering insights and future direction for research.

This paper is organized as follows: [Section 2] provides a literature review of existing studies on sentiment analysis on Amazon product reviews. [Section 3] discusses the studies from the literature review. [Section 4] explains the research methodology used to achieve the paper's objectives. [Section 5] shows the results of the experiments conducted. [Section 6] concludes the paper, and [Section 7] provides suggestions for future research.

2 Literature Review

This section examines existing studies on sentiment analysis in e-commerce, highlighting their contributions, findings, and limitations. Specifically, the studies in this section focus on Amazon product reviews to align with the focus of the current study, as such reviews provide a diverse dataset of buyer opinions. The studies are spread over a six-year period, from 2019 to 2025. This section contains studies that conducted document and sentence-based sentiment analysis, and those that conducted aspect-based sentiment analysis.

2.1 Document and Sentence-Based Sentiment Analysis

[Table 1] lists studies which conducted document and sentence-based sentiment analysis on Amazon product reviews.

Authors	Contributions	Findings	Limitations
[Dey et al. (2020)]	Evaluated SVM and Naïve Bayes in classifying Amazon book reviews into positive and negative.	SVM obtained a higher accuracy and F1-score than Naïve Bayes.	Disregarded neutral reviews and only compared two classifiers.
[Johar and Mubeen (2020)]	Developed a supervised learning model to classify unlabelled Amazon product reviews.	SVM obtained a higher accuracy and F1-score than Naïve Bayes.	Disregarded neutral reviews and only compared two classifiers.
[Urkude et al. (2021)]	Classified Amazon reviews for electronic parts into positive and negative with Naïve Bayes, SVM, Stochastic Gradient Descent	Logistic Regression attained the highest accuracy and F1-score, while Decision Tree had the	Disregarded neutral reviews.

	(SGD), Decision Tree, Logistic Regression and Random Forest.	lowest accuracy and F1-score.	
[Geethangili and Suresh (2022)]	Analyzed Amazon product reviews for word and n-gram for sentiment analysis, with TF-IDF with pruning for feature extraction and Random Forest with feature weights for sentiment classification into positive, negative, and neutral.	Proposed algorithm achieved higher accuracies in all data ranges and all n-grams than Random Forest.	Trigrams needed more processing power and time to generate word vectors.
[Kausar, Fageeri and Soosaimanickam (2023)]	Classified Amazon reviews for Titan men's watches into positive and negative.	Decision Tree attained a higher accuracy than Logistic Regression.	Disregarded neutral reviews.
[Abubaera and Jiddah (2023)]	Classified Amazon product reviews into positive and negative (with and without lemmatization).	Bidirectional Long Short-Term Memory (Bi-LSTM) obtained a higher accuracy and F1-score on lemmatized data.	Disregarded neutral reviews.
[Singh and Khandelwal (2024)]	Employed a convolutional neural network (CNN) to analyze 10,262 Amazon musical instrument reviews.	CNN outclassed other traditional machine learning models.	Ignored neutral reviews.

[Hamza et al. (2024)]	Applied multiple machine learning techniques to analyze Amazon electronic product reviews.	SGD with Logistic Regression achieved the best accuracy among traditional machine learning algorithms, but BERT outperformed all traditional machine learning models.	Only investigated one deep learning model.
[Chenglerayan and Raja (2024)]	Created a BERT-based sentiment analysis model for 23,845 Amazon women's clothing reviews.	BERT attained the highest accuracy and F1-score.	Ignored neutral reviews.
[Ashbaugh and Zhang (2024)]	Compared several machine learning and deep learning models for sentiment analysis on 32,054 Amazon product reviews.	Logistic Regression and Random Forest both achieved the highest accuracy.	Ignored neutral reviews. Positive reviews were overrepresented in the dataset.
[Shetty, H and Mohammed (2025)]	Created an XLNet base-cased model for sentiment analysis on Amazon earphone and PC reviews.	XLNet obtained the highest accuracy and F1-score on both datasets.	XLNet faced difficulties with short reviews, unconventional language, emojis, and product-specific jargon.

Table 1: List of studies which conducted document or sentence-based sentiment analysis on Amazon product reviews.

Firstly, [Dey et al. (2020)] classified 147,000 Amazon book reviews by evaluating Naïve Bayes and SVM. After the preprocessing stage, the authors applied an active learner to dispose of all 3-star reviews. During the feature extraction stage, the authors used TF-IDF and retained nouns with more than 3% frequency for further processing. They refined the collection by selecting the appropriate ones based on opinion words in the reviews. SVM achieved an accuracy and F1 score of 84% and 83.99%, while Naïve Bayes achieved an accuracy and F1 score of 82.88% and 82.66%.

[Johar and Mubeen (2020)] created a supervised learning model to polarize Amazon product reviews. After deleting all 3-star reviews, the authors used part-of-speech (PoS) tagging to create a bag of words consisting of nouns and adjectives and utilized TF-IDF to get the scores of the words. The authors compared Naïve Bayes and

SVM and concluded that SVM achieved a higher accuracy but a lower F1-score, at 82.8% and 82.88%. Naïve Bayes had an accuracy of 82.35% and F1-score of 83.46%.

[Urkude et al. (2021)] conducted sentiment analysis on a balanced dataset with Amazon product reviews of electronic parts. After extracting the features from the data, the authors compared Naïve Bayes, SVM, SGD, Decision Tree, Logistic Regression and Random Forest in classifying the data into positive and negative. Logistic Regression bested the other classifiers, with an accuracy and F1-score of 83.89%, while Decision Tree performed the worst, with an accuracy of 73.3% and an F1-score of 73.27%. Random Forest, Naïve Bayes, SVM and SGD produced accuracies of 83.3%, 82.9%, 82.6% and 78.1%, and F1-scores of 83.28%, 82.89%, 82.69% and 78.09%.

[Geethangili & Suresh (2022)] utilized Random Forest with feature weights to classify Amazon product reviews into positive, negative, and neutral. They used BoW to obtain the frequency of the words, then assigned reviews with 3 stars and above a sentiment rating of 3 (positive), reviews with exactly 3 stars a rating of 2 (neutral), and reviews with less than 3 stars a rating of 1 (negative). The authors assigned weights to the features in the dataset based on Gini index, where the weights were calculated by subtracting the sum of the squares of the probabilities of each class from 1. The authors employed TF-IDF to create word vectors, and pruned words with frequencies outside the range of 0.3 to 1.0. For the classification process, the authors used Random Forest with the Gini index split criteria with 50 trees, a maximum tree depth of 8, and set the voting strategy to majority vote. The authors tested their proposed algorithm with different quantities of data samples, such as 500, 1000, 1500, 2000, and 2500. Their algorithm achieved superior accuracy in all data ranges and all n-grams than the regular Random Forest algorithm, with the highest accuracy being 94.8% at 2500 samples for tri grams. Comparatively, Random Forest achieved a peak accuracy of 92.03% at 2500 samples for trigrams. Both algorithms produced higher accuracies with larger amounts of data as well as longer n-grams. The authors did not disclose the F1-scores of the algorithms.

[Kausar, Fageeri and Soosaimanickam (2023)] classified 4960 Amazon reviews for Titan men's watches into positive and negative. BoW was utilized to extract features from the review text. The authors extracted the top 40 positive and negative review unigrams and bigrams, then split the dataset into training and testing datasets at a 75% to 25% ratio. Logistic Regression and Decision Tree were used to perform sentiment analysis on the data. Logistic Regression had an accuracy of 94%, while Decision Tree had an accuracy of 99%.

[Abubaera and Jiddah (2023)] utilized bi-LSTM to classify Amazon product reviews into positive, negative, and neutral. Because of computing costs, the authors used 2 million reviews out of the 34 million reviews in the dataset. The authors labelled reviews with three stars and above as positive, and those below three stars as negative. The dataset was divided into training and testing datasets at an 80% to 20% ratio. The data was trained in 10 epochs. The authors conducted sentiment analysis in two different scenarios; one where the review text was lemmatized and another where the review text was not lemmatized. Bi-LSTM achieved an accuracy of 96.06% and an F1-score of 90% in the first scenario. In the second scenario, bi-LSTM achieved an accuracy of 93.6% and an F1-score of 87%.

[Singh and Khandelwal (2024)] employed a convolutional neural network (CNN) via SpaCy to analyze 10,262 Amazon musical instrument reviews. Additionally, SpaCy enabled the authors to preprocess the data using tokenization and performing PoS

tagging. They also streamlined the ratings by categorizing 4 and 5-star reviews as positive and the remaining ratings as negative. The data was grouped into training and testing sets at an 80% to 20% ratio. Overall, the model attained an F1-score of 0.79, outperforming other traditional machine learning models such as Logistic Regression, Decision Tree, K-Nearest Neighbors (KNN), SVM and Naïve Bayes. Regardless, the study was limited due to ignoring neutral reviews.

[Hamza et al. (2024)] applied multiple machine learning techniques to analyze Amazon electronic product reviews. The authors preprocessed the data by removing links, numbers and special characters, tokenizing and stemming the text, PoS tagging, removing punctuation, converting text to lowercase and removing stopwords. Random Forest, Multinomial Naïve Bayes, Complement Naïve Bayes, Bernoulli Naïve Bayes, SGD with SVM, SGD with Logistic Regression, and BERT were included in the study. The first six algorithms had two scenarios: one where the data was vectorized with the count vectorizer and the other where it was vectorized with TF-IDF. Out of the traditional machine learning models, SGD with Logistic Regression achieved the best accuracy in both scenarios, at 82.95% and 84.72% respectively. However, BERT outperformed all traditional learning models with an accuracy of 88.93%. One limitation of the study is that it did not explore any deep learning models besides BERT due to resource constraints.

[Chenglerayan and Raja (2024)] created a BERT-based sentiment analysis model for 23,845 Amazon women's clothing reviews. The authors preprocessed the data by removing stopwords and punctuation, removing missing and invalid data, removing duplicate reviews and those with null values, and displaying the text. The authors compared BERT with TF-IDF, Logistic Regression, Random Forest, Naïve Bayes and SVM. BERT obtained the highest accuracy and F1-score, at 96% and 82% respectively. Comparatively, Naïve Bayes obtained the lowest accuracy and F1-score, at 57% and 59% respectively. TF-IDF also achieved an F1-score of 57%. However, the study ignored neutral reviews.

[Ashbaugh and Zhang (2024)] compared several machine learning and deep learning models for sentiment analysis on 32,054 Amazon product reviews. CNN, Recursive Neural Network (RNN), Logistic Regression, Random Forest and Nave Bayes were the models. The authors also assessed the sentiment polarity scores produced by NLTK and TextBlob for further comparison. They initially preprocessed the data by removing punctuation, normalizing the text, removing stopwords, and applying stemming on the text. The data was divided into training and testing datasets at an 80% to 20% ratio. Logistic Regression and Random Forest achieved the highest accuracy, at 99%. TextBlob had the lowest accuracy, at 56%. CNN and RNN achieved accuracies of 98% and 93% respectively, indicating that the deep learning models performed worse than traditional machine learning models. The study was limited by disregarding neutral reviews and its overrepresentation of positive reviews in the dataset.

[Shetty, H and Mohammed (2025)] created an XLNet base-cased model for sentiment analysis on Amazon earphone and PC reviews. There were 14,337 earphone reviews and 30,846 PC reviews. The authors began preprocessing the data by expanding contractions, removing HTML tags, URLs, numbers, and symbols, and converting the text to lowercase. They also tokenized and vectorized the data but did not eliminate stopwords or conduct stemming or lemmatization. The authors compared their XLNet models to several other transformer-based models, such as Distil BERT,

ALBERT, RoBERTa, BERT and ELECTRA. XLNet achieved the highest accuracy and F1-score on both datasets, with 95.2% accuracy and F1-score on the earphone dataset and 95% accuracy and 94.4% F1-score on the PC dataset. However, it faced difficulties with short reviews, unconventional language, emojis, and product-specific jargon. It was also sensitive to subtle linguistic patterns, which caused it to overinterpret neutral or ambiguous reviews.

2.2 Aspect-Based Sentiment Analysis

[Table 2] lists studies which conducted aspect-based sentiment analysis on Amazon product reviews.

Authors	Contributions	Findings	Limitations
[Yang et al. (2019)]	Replaced the conventional attention with a new alternating coattention mechanism that models both the target level and context-level for aspect-based sentiment analysis.	Proposed methods obtained the highest accuracy compared to eight other methods.	Difficult for the proposed methods to learn the complex relationship between words such as the negation modifiers and implicit sentiment phrases.
[Huang and Carley (2019)]	Proposed Target-dependent Graph Attention Network (TD-GAT) for aspect-based sentiment analysis, which explicitly used the dependency relationship among words.	Overall, variants of the proposed model with BERT displayed higher results compared to the variants with Global Vectors for Word Representation (GloVe).	Proposed model did not consider various types of relations in the graph.
[Zhang, Li and Song (2019)]	Proposed to make use of syntactical dependency structures within a sentence and resolved the issue of long-range multi-word dependency for aspect-based sentiment analysis by creating an	ASGCN-directionless (ASGCN-DG) outperformed Aspect-specific Convolutional Neural Network (ASCNN) on all datasets except Restaurant14, but ASGCN-DG and ASGCN-directional (ASGCN-DT)	Proposed model did not exploit edge information of syntactical dependency trees.

	Aspect-specific Graph Convolutional Network (ASGCN) model.	achieved sub-optimal results on the Twitter dataset.	
[Li, Chow and Zhang (2020)]	Proposed semi supervised deep multi-task learning framework (SEML) for aspect mining and aspect sentiment classification sub tasks in aspect-based sentiment analysis.	SEML achieved the highest results for all three tasks on all datasets.	Since SEML directly delivered hidden representations between sub-tasks, it may cause inconsistencies in the results for aspect mining and aspect sentiment classification.
[Al-Ghuribi, Mohd Noah and Tiun (2020)]	Proposed an efficient approach to extract aspects from a domain of interest and aimed to fulfil aspect extraction, aspect weight estimation, inferring aspects' rating and calculating the total review score.	Proposed approach surpassed three other models on three lexicons, in terms of F1-score, accuracy and coverage.	Disregarded neutral reviews. Proposed approach also did not use co-occurrence relations between words during aspect extraction. Some terms that co-occur with some opinion words may not have been extracted.
[Harazeem, Kabir and A (2021)]	Proposed an emoticon-aware aspect-based sentiment analysis model.	Proposed model attained higher precision, recall and accuracy compared to the regular aspect based sentiment analysis model and feature and smiley-based model.	Proposed model requires a trigger to visualize the results.
[Sharma and Kaur (2021)]	Presented a comprehensive statistical comparison on the performance of various deep learning	GAT-BERT achieved the highest average accuracy and average macro F1-score on all datasets, Gated Recurrent Unit (GRU) achieved the	Hyperparameter values were taken from the respective original works for the methods.

	methods on eight datasets for aspect-based sentiment analysis.	lowest average accuracy, and LSTM achieved the lowest average macro-F1 score.	
[Shirahatti, Rajpurohit and Sannakki (2022)]	Introduced a new model using multi-head attention transformation (MHAT) that analyzes words in sentences in parallel and can accurately obtain the global interdependence of those words.	Proposed method accomplished the highest accuracy compared to nine other models.	Did not consider implicit aspects.
[Rachdian, Suryadi and Fransiscus (2022)]	Proposed a method for identifying consumer needs by analyzing product reviews using the lexicon-based method and topic modeling with Non-negative Matrix Factorization (NMF).	Latent Dirichlet Allocation (LDA)'s highest coherence score was 0.3571 with five topics and was more challenging to interpret. The NMF method provided topics that were easier to interpret.	Limitation in topic modeling, where a review may have had more than one topic, and the identified aspect sentiment may have been unrelated to the dominant topic.
[Bellar, Baina and Ballafkih (2024)]	Compared several deep learning models in conducting aspect-based sentiment analysis on 22,641 Amazon product reviews.	The combination of CNN-RNN-Bi-LSTM with Word2Vec obtained the highest accuracy and F1-score.	Spam reviews and fraudulent reviews were not eliminated.
[Xu and Ibrahim (2024)]	Analyzed the performance of aspect-based sentiment analysis in the cross-domain scenario.	Proposed model outperformed the baseline model in all scenarios.	The proposed model encountered difficulties with some terms which were very specific to certain domains.
[Dami and Alimardani (2024)]	Developed a topic modeling approach for aspect-based sentiment analysis	Proposed model attained the highest F1-score on Books,	Authors did not report accuracy scores.

	which could be employed in any product or service.	Electronics and Video Games.	
--	----------------------------------------------------	------------------------------	--

Table 2: List of studies which conducted aspect-based sentiment analysis on Amazon product reviews.

First, [Yang et al. (2019)] proposed Coattention-LSTM, which learns the nonlinear representations of the context and target simultaneously to extract more effective sentiment features from the coattention mechanism, and Coattention-MemNet, which learns the key features from the target and context alternately with an iteration mechanism. The authors assessed their proposed methods on the Restaurant14 and Laptop14 datasets from SemEval (compiled from Amazon), and a Twitter dataset. Compared to eight other methods, Coattention-MemNet produced the highest accuracy on the restaurant dataset, at 79.7%, while Coattention-LSTM achieved the second-highest accuracy at 78.8%. On the laptop and Twitter datasets, Coattention-LSTM obtained the highest accuracy, at 73.5% and 71.5% respectively. Meanwhile, Coattention-MemNet obtained accuracies of 72.9% and 70.5% on the two datasets. The authors did not provide the macro F1-scores of the models. A notable limitation of the proposed methods is that it is difficult for them to learn the complex relationship between words such as the negation modifiers and implicit sentiment phrases.

[Huang and Carley (2019)] proposed a TD-GAT for aspect-based sentiment analysis, which explicitly used target information by using a LSTM to model the aspect target dependency across layers and overcome noisy information. The authors used the laptop and restaurant review datasets from SemEval to evaluate their model. They used the Stanford dependency parser to obtain the dependency graphs and tried two embedding methods: 300-dimensional GloVe embeddings and BERT representations. TD-GAT-BERT with five layers obtained the highest accuracy on the laptop review dataset, at 80.1%, while TD-GAT-BERT with four layers obtained the highest accuracy on the restaurant review dataset, at 83%. The authors did not provide the macro F1-scores of the models. Overall, the variants of the proposed model with BERT performed better than the variants with GloVe. However, the proposed model only used the dependency graph and did not consider various types of relations in the graph.

[Zhang, Li and Song (2019)] proposed to utilize syntactical dependency structures within a sentence and fixed the issue of long-range multi-word dependency for aspect-based sentiment analysis. From this, the authors proposed an ASGCN model, and two variants on dependency graphs which were directionless (ASGCN-DG) and directional (ASGCN-DT). The variants were performed in a multi-layer manner, on the top of a bi-LSTM. The authors evaluated their proposed model on five datasets; Twitter, Laptop14, Restaurant14, Restaurant15 and Restaurant16, and compared it to five other models. They initialized the word embeddings using 300-dimensional GloVe vectors. The results were obtained by averaging the accuracy and macro F1-score of three runs. ASGCN-DG achieved higher results than the other models on the Laptop14 and Restaurant15 datasets from SemEval, achieving an accuracy of 75.55% and a macro F1-score of 71.05% on the Laptop14 dataset, and an accuracy of 79.89% and a macro F1-score of 61.89% on the Restaurant15 dataset. Transformation Network-Lossless Forwarding (TNET-LF) achieved the highest results on the Twitter and Restaurant16 datasets, with an accuracy of 72.98% and a macro F1-score of 71.43 on the Twitter

dataset, and an accuracy of 89.07% and a macro F1-score of 70.43% on the Restaurant16 dataset. ASCNN achieved the highest results on the Restaurant14 dataset, with an accuracy of 81.73% and a macro F1-score of 73.1%. Other than that, ASGCN-DG and ASGCN-DT achieved sub-optimal results on the Twitter dataset due to the sentences in the dataset containing grammatical errors.

[Li, Chow and Zhang (2020)] proposed SEML for aspect mining and aspect sentiment classification sub-tasks in aspect-based sentiment analysis which applied CVT to utilize unlabelled reviews, thus improving the representation learning within a unified end-to-end architecture. SEML consisted of four main components; representation learning, aspect mining, aspect sentiment classification and CVT, and used three stacked bidirectional recurrent neural layers with moving-window attention mechanism within the Gated Recurrent Unit (MAGRU) to build the shared contextualized representation learning component for aspect mining and aspect sentiment classification. Because each layer used the outputs from the layer below it as inputs, the proposed framework enabled multitask learning and interaction between the different sub-tasks to improve aspect extraction and aspect sentiment prediction. To enable semi-supervised learning, the proposed framework was trained on both supervised and non-supervised reviews using CVT. The authors employed pre-trained GloVe and refined the sentiment vectors to initialize the word embeddings. SEML achieved the highest results for the aspect mining, aspect sentiment classification and aspect-based sentiment analysis on all datasets. For aspect mining, SEML obtained an F1-score of 83.37% on the laptop dataset.

[Al-Ghuribi, Mohd Noah and Tiun (2020)] proposed an efficient approach which was divided into aspect extraction, aspect weight estimation, aspect rating inference, and the calculation of the review score. During the aspect extraction stage, the authors extracted nouns and noun-adjective pairs from the reviews and calculated the frequency of the nouns using a combination of a frequency-based approach and a syntactic relation-based approach. The nouns and their frequencies were then stored in blocks to reduce the number of searching comparisons in the datasets. The second process in the aspect extraction stage was the creation of the main dictionary, which contained every word and the number of times it appeared as a noun, the number of times it appeared with the adjectives, and the number of times it appeared with the comparative or superlative adjectives. The third process in the stage was the generation of the main aspects and core terms. In the aspect weight estimation stage, the authors utilized TF-IDF, and two variants proposed by [Zhu, Wang and Zou (2016)] and [Ngoc, Thu and Nguyen (2019)]. During the aspect rating inference stage, the rating of each aspect was calculated by extracting the sentiment words for the aspect using the syntactic parser and the named entity recognition (NER) provided by the Java Stanford-CoreNLP library and assigning a score to the sentiment words. The authors used the results of the other three tasks to calculate the score of the reviews. The authors evaluated their approach on 1,500,000 book and 1,300,000 movie reviews from Amazon, and 1,000,000 restaurant reviews from Yelp. The reviews were divided into five classes, corresponding to the value of the rating assigned to each review, though the authors disregarded neutral reviews. The proposed approach surpassed three other models on three lexicons tested by the authors, in terms of F1-score, accuracy and coverage. However, it did not employ co-occurrence relations between words during aspect extraction. Due to this, some terms that co-occur with some opinion words may not have been extracted.

[Harazeem, Kabir & A (2021)] proposed an emoticon-aware aspect-based sentiment analysis model. They built an emoticon dictionary so the model could automatically detect the emoticons and calculate the score of the emoticons in the reviews. The emoticons were given one of three labels: positive, negative, or neutral. Positive emoticons were assigned a score of 1, negative emoticons were assigned a score of -1, and neutral emoticons were assigned a score of 0. The final output of the model was the aspects and their associated polarity. The authors evaluated their proposed model using a dataset of 2085 iPhone reviews from Amazon and compared it to a non-emoticon-aware aspect-based sentiment analysis model and a feature and smiley-based model. The proposed model obtained a higher precision, recall and accuracy than the other models, at 88.1%, 84.6% and 88.5% respectively. The non-emoticon-aware model achieved 62.6% precision, 62.4% recall and 62.1% accuracy, and the feature and smiley-based model achieved 63.45% precision and 65.73% recall, but the authors did not evaluate its accuracy.

[Sharma and Kaur (2021)] presented a comprehensive statistical comparison on the performance of 35 deep learning methods on eight datasets: Restaurant14, Laptop14, Restaurant15, Restaurant16, Twitter, Sentihood, Mitchell, and Multi-Aspect Multi-Sentiment (MAMS), for aspect-based sentiment analysis. For the non-BERT-based methods, the authors used GloVe embeddings. For the BERT-based methods, the authors used the pre-trained BERT embeddings. GAT-BERT performed the highest on all datasets, with the highest average accuracy and average macro F1-score, at 84.78% and 73.34% respectively. GRU performed the lowest in terms of average accuracy, and LSTM performed the lowest in terms of average macro-F1 score. Overall, the top 10 highest performing methods were GAT-BERT, BERT with sentence pair classification (BERT-SPC), Relational Graph Attention Network (RGAT), CapsNet, Attention Encoder Network-BERT (AEN BERT), ASGCN, aspect specific Tree Convolution Network (ASTCN), ASCNN, TNET and TD LSTM, all of which achieved consistent results for all datasets, except TNET, which performed poorly on the Restaurant15 dataset.

[Shirahatti, Rajpurohit and Sannakki (2022)] introduced a new model using MHAT that analyzes words in sentences in parallel and can accurately obtain the global interdependence of those words. The authors used the Laptop14 dataset from SemEval in their study, which contained laptop reviews from Amazon. They employed GloVe for word embedding and applied PoS tagging to each word in the dataset. They then provided their model with the relative positions of the token through position encoding. After that, the authors utilized multi-head attention to capture the context in which the aspects were mentioned and consequently trained and tested their model. The proposed method accomplished the highest accuracy compared to nine other models. The model did not consider implicit aspects in the study.

[Rachdian, Suryadi and Fransiscus (2022)] proposed a methodology to detect consumer needs by analyzing product reviews using the lexicon-based method and topic modeling with NMF. They intended to transform the results of aspect-based sentiment analysis into consumer needs with its priority ranking. The dataset utilized in the study included laptop reviews from Amazon. The authors used TF-IDF for feature extraction, and the document-term matrix from the feature extraction stage was used as the input for topic modeling. After the topics were defined, aspect-based sentiment analysis was conducted. Once the aspects and opinions were identified, the reviews were scored, and the intensity of the aspect sentiments were calculated. Finally,

the consumer needs were ranked based on importance. LDA's highest coherence score was 0.3571 with five topics and was more difficult to interpret. In comparison, the NMF method provided topics that were easier to interpret. The authors' proposed method had limitations, particularly in the topic modeling, where a review may have had more than one topic, and the identified aspect sentiment may not have related to the dominant topic. The proposed method was also unable to handle sarcasm, opinion words with a wide meaning, and substitute words.

[Bellar, Baina and Ballafkih (2024)] compared several deep learning and machine learning models in conducting aspect-based sentiment analysis on 22,641 Amazon product reviews. They preprocessed the data by changing the text to lowercase, deleting extra spaces, numbers, punctuation and stopwords, and tokenizing and lemmatizing the text. They compared some models in three-class and five-class scenarios, while others were only compared in the three-class scenario. In the three-class scenario, 4 and 5-star reviews were categorized as positive, 3-star reviews were categorized as neutral and 1 and 2-star reviews were categorized as negative. The models performed better with fewer classes. In the three-class scenario, the combination of CNN-RNN-Bi-LSTM with Word2Vec accomplished an accuracy and F1-score of 96.2% and 91.3% respectively, which were the best results. In contrast, Random Forest achieved the lowest accuracy and F1-score at 55.02% and 24.8% respectively. Despite the good results produced, the authors stated that spam reviews and fraudulent reviews were not eliminated, which could have affected the accuracy of the models.

[Xu and Ibrahim (2024)] analyzed the performance of aspect-based sentiment analysis in the cross-domain scenario where models trained on the review data of one domain were applied to another domain. The data was collected from several e-commerce platforms such as Amazon, eBay and Alibaba, and encompassed several categories, including electronics, fashion, and home appliances. The authors preprocessed the data by eliminating HTML tags, URLs and special characters, tokenizing, stemming and lemmatizing the text, and removing stopwords. After that, the authors extracted aspects from the reviews using both rule-based and machine learning techniques. They then classified the aspect terms as positive, negative or neutral. The models were fine-tuned to handle domain-specific variations of conveying sentiment. The authors employed transfer learning to enable the models to adapt to reviews from different domains. The authors conducted experiments in three scenarios: traditional sentiment analysis, single-domain aspect-based sentiment analysis, and cross-domain aspect-based sentiment analysis. The proposed model outperformed the baseline model in all instances. The proposed model trained on the electronics dataset and tested on the fashion dataset achieved an accuracy of 0.74 and F1-score of 0.77. When trained on the fashion dataset and tested on the home appliances dataset, it achieved an accuracy of 0.71 and an F1-score of 0.74. The results indicate that the model can generalize across domains, although it did face difficulties with some terms which were very specific to certain domains.

[Dami and Alimardani (2024)] developed a topic modeling approach for aspect-based sentiment analysis which could be used for any product or service. They preprocessed the data by extracting features from the text with BoW. Then, they utilized LDA to extract topics from the reviews. After that, they combined their topical modeling approach with SVM to perform aspect-based sentiment analysis on Amazon product reviews from different categories such as Books, Electronics, Cell Phones and Accessories, Health and Personal Care and Video Games. The proposed model

accomplished the highest F1-score on Books, Electronics and Video Games at 97.49% on Books, 99.1% on Electronics, and 79.01% on Video Games. On Cell Phones and Accessories, it accomplished the lowest F1-score at 90.79%. In contrast, bagging achieved the highest F1-score at 98.33%. On Health and Personal Care, the proposed model achieved an F1-score of 51.05, with the highest being 53.57% by bagging and lowest being 46.4% by LSTM-CRF. The authors did not report the accuracy of the models.

3 Discussion

The studies analyzed in the previous section reveal that there are common challenges or research gaps in sentiment analysis that could be overcome by future studies.

Firstly, most studies conducted sentiment analysis on product reviews on electronic products, particularly laptops. Product reviews from other categories are still important because different product categories may have different characteristics which can impact a buyer's sentiment, and limiting sentiment analysis to one category may limit the generalizability of the results. Reviews for a particular product can contain domain specific terminology not used in reviews for other types of products. Conducting sentiment analysis on reviews from a variety of product categories can lead to a more detailed perception of similarities and differences in reviews and the aspects that affect them.

Secondly, many studies disregarded neutral reviews, mainly those which conducted document and sentence-based sentiment analysis. Such reviews are usually more difficult to classify as the sentiments expressed are less explicit than those in positive and negative reviews. Nevertheless, the reviews still contain useful information that should not be ignored. For example, a neutral review can contain a buyer's opinion on aspects which are not positive or negative but may impact other buyers' purchase decisions. In certain cases, a high quantity of neutral reviews can denote mixed opinions on a product, which implies room for improvement for sellers or manufacturers. It is important for studies on sentiment analysis to recognize biases that can arise from excluding neutral reviews. Future studies should use methodologies which include neutral reviews in sentiment analysis.

Thirdly, some studies on document and sentence-based sentiment analysis only compared two classifiers. Although these comparisons can provide basic insights into the classifiers' performance, they do not offer a comprehensive understanding of their capabilities. Future work should compare more classifiers to discern which classifiers achieve better performance in different scenarios and how to overcome the weaknesses of the classifiers with poor results. Comparing more classifiers can also allow researchers to determine which types display better generalizability in different domains, which is important for real-world applications.

4 Research Methodology

This section outlines the methodology used to achieve the set objectives systematically.

4.1 Data Collection Method

The Amazon product reviews used in this study were compiled by [Ni, Li and McAuley (2019)]. This study made use of the 5-core versions of the datasets, where each product had at least five reviews each. This study used two datasets, which were Industrial and Scientific and Luxury Beauty. These datasets were chosen as they had smaller file sizes than the other datasets due to having fewer reviews, while also maintaining diversity in terms of the product categories. This study can contribute to the field through its focus on other product categories that are less explored in previous studies. The number of reviews per dataset prior to any data manipulation is shown in Table 3.

No.	Dataset	No. of Reviews
1	Industrial and Scientific	77,071
2	Luxury Beauty	34,278
Total		111,349

Table 3: Number of reviews per dataset.

4.2 Research Framework

4.2.1 Phase 1

[Figure 1] shows Phase 1 of the research framework.

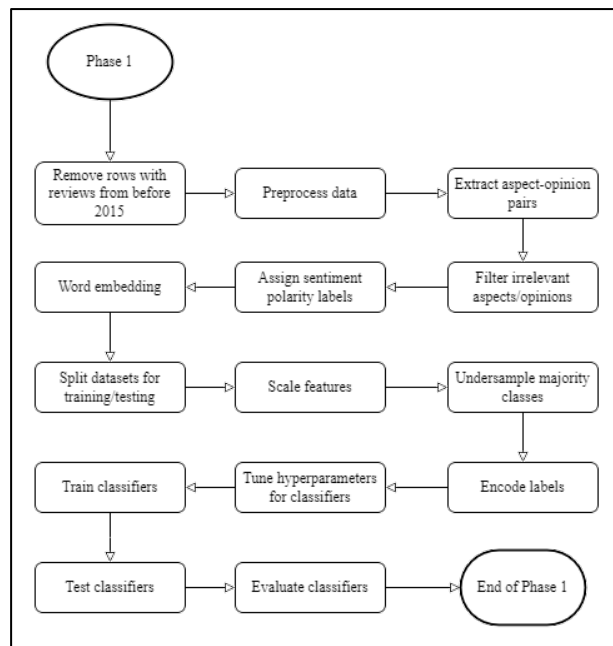


Figure 1: Research Framework (Phase 1)

In the first stage of Phase 1, all rows with reviews from before 2015 were removed. Pandas was used to convert the dates in the reviewTime column to the MM DD, YY format. Then, the max_date variable was defined as January 1, 2015. After that, filtered_dataset was defined to only keep all rows where reviewTime exceeded max_date. Since this caused many rows to be dropped, the indices of the dataset were reset to ensure they would remain consecutive.

To begin the preprocessing stage, some columns which lacked meaningful information for sentiment analysis were dropped from the dataset, including Vote, Verified, Style, Overall, Summary, Unix Review Time, Image, Reviewer Name and Reviewer ID. The remaining columns are shown in [Table 4].

Data Column	Purpose
ASIN	The product for which a review is written.
Review Text	The main body of the review.
Review Time	The time the review was written, in raw format.

Table 4: Remaining data columns in Amazon product review datasets.

After that, rows with null values were dropped so that only complete data would be processed. Preprocessing was applied to the reviewText column using the Re and SpaCy libraries to replace special characters, HTML code and numbers with spaces, expand contractions, convert the text to lowercase and remove stopwords. Hence, the data was sanitized so that it could be read and processed by the algorithms.

The aspects and opinions from the review text were subsequently extracted using SpaCy. It detected the PoS of each word and traversed the dependency tree of each review to capture relationships between word tokens. For example, if a token was categorized as a noun, it was identified as an aspect, and if another token positioned nearby was an adjectival modifier for the noun, it was identified as an opinion. The two tokens were extracted as part of an aspect-opinion dictionary pair. If an aspect was identified without an associated opinion or vice versa, the dictionary was removed. If a review had no aspect-opinion dictionaries, it was dropped from the dataset. A new column was added to each row in the datasets to store the aspect-opinion dictionaries, and the indices were again reset to ensure they would be consecutive. [Figure 2] shows a review snippet from the Industrial and Scientific dataset after aspect and opinion extraction.

```
{
  "reviewTime":1495238400000,
  "asin":"B0000223SK",
  "reviewText":"good product",
  "aspects_opinions":[
    {
      "aspect":"product",
      "opinion":"good"
    }
  ]
},
```

Figure 2: Review snippet from Industrial and Scientific dataset after aspect and opinion extraction.

However, because SpaCy's extraction works based on rules, it may have extracted words that did not contribute to the sentiment of the review. Thus, aspect-opinion filtering was done using the Counter library to count the frequency of each aspect and opinion. If an aspect-opinion pair appeared less than 10 times throughout the whole dataset, it was deemed irrelevant and the dictionary it is from was removed.

After that, sentiment polarity labels were assigned to the aspect-opinion pairs, based on the scoring of the opinion assigned by the TextBlob library. If an opinion had a score higher than 0, it was labelled positive. If the score was lower than 0, it was labelled negative. If the score was exactly 0, it was labelled neutral. [Figure 3] shows a snippet of a review from the Industrial and Scientific dataset after sentiment polarity labelling.

```
{
  "reviewTime":1495238400000,
  "asin":"B0000223SK",
  "reviewText":"good product",
  "aspects_opinions":[
    {
      "aspect":"product",
      "opinion":"good",
      "sentiment":"positive"
    }
  ]
},
```

Figure 3: Review snippet from Industrial and Scientific dataset after sentiment polarity labelling.

After that, word embedding was performed. The specific version of GloVe used was the Common Crawl which has the parameters of 840 billion tokens and 300 dimensions [Pennington, Socher and Manning (2014)]. To load GloVe, an embeddings dictionary was initialized, and each word in the GloVe file and its associated vector were extracted and added to the embeddings dictionary as pairs. Lines from the file that could not be parsed were skipped.

After the GloVe model was loaded, the average embeddings of the aspect and opinion from each dictionary were calculated and combined. Using combined embeddings allowed the relationship between an aspect and opinion to be understood by the classifiers. The combined embeddings were added to each aspect-opinion dictionary. Also, during this stage, the datasets were arranged so the reviews were stored as dictionaries under their parent product. This allowed for a more comprehensive view of the hierarchy between a product, its reviews, and the reviews' aspect-opinion dictionaries. Since the original review IDs were not provided in the dataset, new review IDs were assigned to each review on a per-product basis.

Following the work performed by other researchers such as [Bansal and Srivastava (2018)] and [Alharbi et al. (2021)], the datasets were divided at a ratio of 80% training data to 20% testing data using Scikit-Learn's `train_test_split` function. The next step was to perform feature scaling using Scikit-Learn's `StandardScaler`, which standardizes features by removing the mean and scaling to unit variance [Scikit Learn Developers (2024d)]. Feature scaling, also known as normalization, helps to standardize the scale of the data, which leads to more accurate results [Nkikabahizi, Cheruiyot and Kibe (2022)]. After the scaler was set to fit the training data, the training and testing data were scaled, and the datasets were updated to store the scaled data instead of the original data.

The majority classes in the training datasets were then undersampled using Imbalanced-Learn's `RandomUnderSampler` [Imbalanced-Learn Developers (2024)]. Maintaining the class imbalance would lead to the classifiers being biased towards the majority classes, resulting in poor generalization and overfitting. Undersampling reduces the amount of data in the majority classes to create a more balanced class distribution in the datasets.

To prepare for the classifiers to be trained, tested and evaluated, label encoding was applied to the class labels using Scikit-Learn's `LabelEncoder`. The "positive", "negative" and "neutral" labels were assigned a value of 0, 1 and 2 respectively. Following that, the hyperparameters for Decision Tree and SVM were tuned using Scikit-Learn's `RandomizedSearchCV` function, which allows randomized combinations of hyperparameter values to be tested on the data using cross-validation. This study used the macro F1-score as the main scoring metric, following in the footsteps of other researchers such as [Zhang, Li and Song (2019)], [Danistan et al. (2020)], and [Li, Chow and Zhang (2020)]. 10 combinations of hyperparameters were randomly tested, and each combination was cross-validated five times, which is the default number of folds [Scikit-Learn Developers, (2024b)]. For SVM, the class weight was set to "balanced" to ensure that each class would be treated equally and reduce bias to the majority classes. Hyperparameter tuning was not performed on Naïve Bayes or Random Forest. Naïve Bayes has very few hyperparameters compared to the other three classifiers in this study [Scikit-Learn Developers, (2024c)]. As such, its default hyperparameters were maintained. As for Random Forest, since it is an ensemble model made up of several Decision Trees, the hyperparameters obtained from the

RandomizedSearchCV for Decision Tree were used, in addition to Random Forest's default hyperparameter of having 100 Decision Trees. The hyperparameter grids for Decision Tree and SVM are shown in [Table 5] and [Table 6].

Hyperparameter	Values	Description
max_depth	3, 5, 7	Maximum depth of the tree
min_samples_split	2, 5, 10	Minimum number of samples required to split an internal node
min_samples_leaf	1, 2, 5	Minimum number of samples required to be at a leaf node
max_features	sqrt (square root), log2	Number of features to consider when looking for the best split
criterion	gini, entropy	Measures the quality of a split

Table 5: Hyperparameter grid for Decision Tree.

Hyperparameter	Values	Description
c	0.1, 1	Regularization parameter
kernel	linear, rbf	The kernel type
gamma	1, 0.1, 10	Kernel coefficient

Table 6: Hyperparameter grid for SVM.

In the final stage of Phase 1, Gaussian Naïve Bayes, Decision Tree, Random Forest and SVM were trained on the combined embeddings of the aspect-opinion dictionaries in the training datasets. After the training stage, the classifiers were tested on the testing datasets. The results produced by the classifiers in this stage were mainly evaluated using the precision, recall, accuracy, and macro F1-score metrics.

4.2.2 Phase 2

[Figure 4] shows Phase 2 of the research framework.

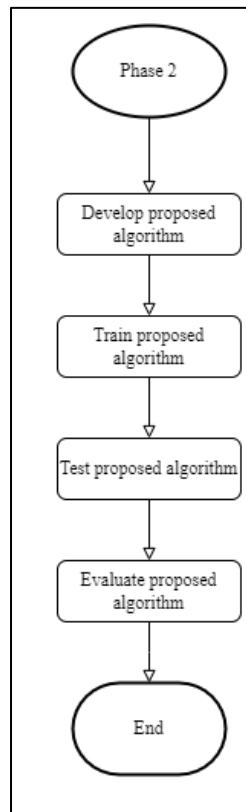


Figure 4: Research Framework (Phase 2)

In Phase 2, the top two classifiers with the best average macro F1-scores from Phase 1 were utilized to develop SVHA. The main goal behind using a hybrid algorithm is to make use of the predictions of different base classifiers. By doing so, the proposed algorithm can compensate for the weaknesses of each base classifier and produce more accurate results [Kacem et al. (2020)].

The proposed algorithm in this study follows an ensemble-based architecture. It was implemented using a voting classifier, which combines the outputs of base classifiers to produce the final output. The type of voting used is soft voting, which produces the final prediction by selecting the class with the highest average sum of probabilities from the base classifiers [Scikit-Learn Developers (2024b)]. Hence, soft voting allows SVHA to consider the confidence level of each base classifier’s predictions rather than strictly classifying an aspect opinion pair as positive, negative or neutral. The proposed algorithm was trained and tested on the same datasets from Phase 1, and evaluated using the precision, recall, accuracy, and macro F1-score metrics. As Random Forest and SVM achieved the best average macro F1-score across the two datasets, they were used as the base classifiers. The hyperparameters for those classifiers remain the same as in Phase 1, with the addition of the “probability” parameter in SVM set to “True”.

5 Results

This section presents the results of the aspect-based sentiment analysis produced by Naïve Bayes, Decision Tree, Random Forest, SVM and SVHA. [Table 7] shows a summary of results obtained by all five algorithms in this study.

Classifier	Dataset	No. of Samples	Accuracy	Macro F1-score
Naïve Bayes	Industrial and Scientific	64,624	75%	74%
	Luxury Beauty	61,258	76%	75%
Decision Tree	Industrial and Scientific	64,624	87%	86%
	Luxury Beauty	61,258	83%	82%
Random Forest	Industrial and Scientific	64,624	98%	98%
	Luxury Beauty	61,258	98%	98%
SVM	Industrial and Scientific	64,624	99%	98%
	Luxury Beauty	61,258	99%	99%
SVHA	Industrial and Scientific	64,624	99%	99%
	Luxury Beauty	61,258	99%	99%

Table 7: Summary of results.

The ranking of the algorithms in terms of highest overall accuracy, macro F1-score and number of correctly classified aspect-opinion pairs is as follows: SVHA ranked first, Random Forest ranked second, SVM ranked third, Decision Tree ranked fourth and Naïve Bayes ranked fifth. Naïve Bayes' results were substantially lower than those obtained by the other four algorithms, as it obtained accuracies and macro F1-scores ranging from 75% to 76%. In comparison, Decision Tree attained accuracies and macro F1-scores ranging from 82% to 87%, whereas Random Forest, SVM and SVHA obtained accuracies and macro F1-scores ranging from 98% to 99%.

Compared to Random Forest, which obtained the second-highest results overall, SVHA was an improvement in both datasets as well as overall, as it was able to correctly classify a higher number of aspect-opinion pairs. However, on the Luxury Beauty dataset and on average, Random Forest correctly classified a higher number of neutral pairs than SVHA. Naïve Bayes achieved better results on the Luxury Beauty dataset than Industrial and Scientific, while the reverse is true for Decision Tree and SVM. The results of Random Forest and SVHA on both datasets were roughly equal.

A significant issue with SVHA is its lack of efficiency. Despite the improved results it obtained compared to the other four algorithms, it took around 16 hours to

train and test the two datasets. This is likely due to one of its base classifiers being SVM, which has previously been reported to take a long time to train [Aboobaker and Ilavarasan (2020)] because the training kernel matrix increases in quadratic form as the dataset gets bigger [Sachin and Kumar (2022)].

There are some considerations regarding the results obtained by the five algorithms used in this study that should be acknowledged, which are mostly related to the earlier steps in the implementation of the research methodology as outlined in [Section 3]. Firstly, the aspect and opinion extraction by SpaCy was not entirely accurate. In some cases, an aspect and opinion extracted as a pair were not related in the original review. The extraction process also extracted some aspects and opinions that did not contribute to the sentiment of the review. Secondly, due to the method for aspect and opinion filtering being based on frequency, some irrelevant aspect-opinion pairs were still retained because it met the minimum required number of occurrences. Some potentially relevant pairs were also removed due to not meeting the minimum required number of occurrences. Lastly, TextBlob may have assigned incorrect sentiment scores to some aspect-opinion pairs, which would have affected the labelling.

As for the effects of the review lengths and language complexities, it was found that the aspect and opinion extraction performed better on shorter reviews with less complex language, as SpaCy efficiently extracted the relevant aspects and their associated opinions. However, in some cases, it attributed some opinions to the wrong aspects, and extracted aspects and opinions which did not contribute to the sentiment of the review. These issues were exacerbated on longer reviews as well as those with more complex language.

6 Conclusion

As mentioned previously, there is no easy way for sellers and buyers to understand and analyse other buyers' sentiments on products being sold on many e-commerce platforms. It is time-consuming to read all reviews [Jiang et al (2021)] and identifying aspects of a product that are mentioned positively or negatively in a review is difficult without reading it manually or using third-party tools. It is still feasible to read a small quantity of reviews, but less so when the total number of reviews is high as it can lead to information overload, which can negatively affect buyers' decision-making process [Hu and Krishen (2019)]. This study addressed this with three research objectives as outlined in [Section 1].

The steps taken to complete this study have proven that aspect-based sentiment analysis can be leveraged to improve the e-commerce shopping experience for buyers by providing a summary of the most frequent aspects mentioned in the reviews for a product, along with their associated polarities, providing vital information that can influence buyers' purchasing decisions while mitigating information overload. This was mainly shown via the development and implementation of SVHA.

The work done in this study can benefit several parties. First, the extensive literature review conducted in [Section 2] can serve as a starting point for researchers from academic and business landscapes to conduct more research on e-commerce, product reviews, machine learning, NLP and sentiment analysis, based on the challenges and research gaps discovered. Next, NLP practitioners can apply and explore the techniques used in the implementation of the research methodology to

conduct more thorough experiments on aspect-based sentiment analysis, considering the suggestions for future work from the previous section. The techniques employed in this study can be extended into other domains that require the analysis of user-generated content such as social media.

7 Future Work

This study has several limitations. Firstly, although SpaCy was able to extract aspects and opinions from the reviews, there were cases where it extracted an aspect and opinion which were unrelated in the original review or extracted irrelevant aspects and opinions. SpaCy's reliance on syntactic rules can cause difficulties with complex language structures. Future work can alleviate this issue by utilizing a more sophisticated aspect and opinion extraction method which does not rely solely on rules to understand the relationship between aspects and opinions, such as transformer-based models. This can ensure that the extracted aspects and opinions are related to each other and relevant to the sentiment of the review.

Secondly, the aspect and opinion extraction method in this study only extracted individual words as aspects and opinions. Although this covered a wide range of aspects mentioned and opinions expressed, there may have been multi-word aspects and opinions which contributed to the sentiment of the review which were not extracted. Therefore, some contextual information may have been lost, resulting in an incomplete view of a buyer's sentiment towards a product. Thus, future work can use an extraction method which considers both single-word and multi-word aspects, such as noun phrase extraction.

Furthermore, the aspect and opinion extraction method in this study was unable to detect implicit aspects. As implicit aspects are only implied and not stated, they are difficult to detect [Ganganwar and Rajalakshmi (2019)]. This is because semantic understanding is required to detect them, by determining the context from the review text. It is important to identify implicit aspects as many sentiments expressed in reviews are not overly expressed. Future work can attempt to develop a method which can detect and extract both implicit and explicit aspects to provide a more accurate view of the sentiments expressed in product reviews. This can be accomplished by employing advanced NLU techniques which utilize contextual word embeddings such as those in transformer-based models. Alternatively, methods that incorporate domain-specific knowledge can also be employed to identify implicit aspects that are common to specific product categories.

Moreover, the aspect filtering method in this study worked based on frequency. Consequently, some aspect-opinion pairs were still retained because they met the minimum required number of instances, regardless of their relevance. Conversely, some potentially relevant pairs were removed due to not meeting the minimum required number of instances. In future studies, this can be rectified by using a filtering method that uses context to determine relevance, rather than frequency. This can be done by using semantic filtering, which can evaluate the relevance of aspect-opinion pairs to their parent reviews.

In addition, this study removed reviews where no aspects were detected. Since the focus of this study was on aspect-based sentiment analysis, reviews with no aspects were considered anomalies. Such reviews can still provide insights into buyers'

sentiments, particularly when they express opinions about a product in general rather than specific aspects. Future work can manage these anomalies by assigning a placeholder aspect and opinion based on the overall rating of the review.

As this study has some limitations in terms of the data used, it can be extended in the future to extract reviews for products from Amazon in real-time, rather than relying on outdated data obtained from datasets. Therefore, the results of the aspect-based sentiment analysis would be more accurate to the current perception of the products on Amazon.

Acknowledgements

This work was supported by Taylor's University through its Taylor's Research Scholarship Programme.

References

- [Aboobaker and Ilavarasan (2020)] Aboobaker, J. and Ilavarasan, E. (2020): "A Survey on Sarcasm Detection Approaches"; *Indian Journal of Computer Science and Engineering*, 11, 6 (2020), 751–771.
- [Abubaera and Jiddah (2023)] Abubaera, M.M. and Jiddah, S.M.: "Natural Language Processing and Bi-Directional LSTM for Sentiment Analysis"; *International Journal of Computer Applications*, 185, 27 (2023), 31–35.
- [Al-Ghuribi, Mohd Noah and Tiun (2020)] Al-Ghuribi, S.M., Mohd Noah, S.A. and Tiun, S.: "Unsupervised Semantic Approach of Aspect-Based Sentiment Analysis for Large-Scale User Reviews"; *IEEE Access*, 8 (2020), 218592–218613.
- [Alharbi et al. (2021)] Alharbi, N.M., Alghamdi, N.S., Alkhamash, E.H. and Al Amri, J.F.: "Evaluation of Sentiment Analysis via Word Embedding and RNN Variants for Amazon Online Reviews"; *Mathematical Problems in Engineering*, 2021 (2021), 1–10.
- [Ashbaugh and Zhang (2024)] Ashbaugh, L. and Zhang, Y.; "A Comparative Study of Sentiment Analysis on Customer Reviews Using Machine Learning and Deep Learning. *Computers*"; 13, 12 (2024), 1–16.
- [Avlani (2020)] Avlani, A. K.: "E-Commerce-An Evaluation of Evolution"; *International Journal of Research in all Subjects in Multi Languages*, 8, 10 (2020), 27 - 31.
- [Bansal and Srivastava (2018)] Bansal, B. and Srivastava, S.: "Sentiment classification of online consumer reviews using word vector representations"; *Procedia Computer Science*, 132 (2018), 1147–1153.
- [Bellar, Baina and Ballafkih (2024)] Bellar, O., Baina, A. and Ballafkih, M.; "Sentiment Analysis: Predicting Product Reviews for E-Commerce Recommendations Using Deep Learning and Transformers"; *Mathematics*, 12, 15 (2024), 1–21.
- [Chenglerayen and Raja (2024)] Chenglerayen, K. and Raja, S.R.; "From Reviews to Results: Leveraging Amazon Feedback for Product Evolution"; *International Journal of Scientific Research in Science, Engineering and Technology*, 11, 6 (2024), 321–328.
- [Dami and Alimardani (2024)] Dami, S. and Alimardani, R.; "An Aspect-Level Sentiment Analysis Based on LDA Topic Modeling"; *Journal of Information Systems and Telecommunication*, 12, 2 (2024), 117–126.

- [Danistan et al. (2020)] Danistan, R., Arunakirinathan, T., Sivarajah, A., Mehendran, Y., & Ekanayake, J.; “Aspect Based User Reviews Classification”; *Instrumentation*, 7, 2 (2020), 9–19.
- [Dey et al. (2020)] Dey, S., Wasif, S., Tonmoy, D.S., Sultana, S., Sarkar, J. and Dey, M.: “A Comparative Study of Support Vector Machine and Naive Bayes Classifier for 24 Sentiment Analysis on Amazon Product Reviews”; *IEEE Xplore* (2020), 217–220.
- [Ganganwar and Rajalakshmi (2019)] Ganganwar, V. and Rajalakshmi, R.; “Implicit Aspect Extraction for Sentiment Analysis: A Survey of Recent Approaches”, *Procedia Computer Science*, 165 (2019), 485–491.
- [Geethangili and Suresh (2022)] Geethangili, D. and Suresh, P.: “Machine Learning Approach based Sentiment Analysis, Classification: An Application of Natural Language Processing”; *Mathematical Statistician and Engineering Applications*, 71, 4 (2022), 773 - 786.
- [Hamza et al. (2024)] Hamza, A., Majeed, K.B., Muhammad, R. and Jaffar, A.; “An Integrated Approach for Amazon Electronic Products Reviews by Using Sentiment Analysis”; *Bulletin of Business and Economics*, 13, 2 (2024), 142–153.
- [Harazeem, Kabir and A (2021)] Harazeem, A.O., Kabir, U. and A, K.H.: “Emoticon Aware Aspect Based Sentiment Analysis of Online Product Review”; *Dutse Journal of Pure and Applied Sciences (DUJOPAS)*, 7, 2a (2021), 166–179.
- [Hu and Krishen (2019)] Hu, H. and Krishen, A.S.: “When is enough, enough? Investigating product reviews and information overload from a consumer empowerment perspective”; *Journal of Business Research*, 100 (2019), 27–37.
- [Huang and Carley (2019)] Huang, B. and Carley, K.M.: “Syntax-Aware Aspect Level Sentiment Classification with Graph Attention Networks”; *ArXiv* (2019), 1–9.
- [Imbalanced-Learn Developers (2024)] Imbalanced-Learn Developers; “RandomUnderSampler”; *Imbalanced Learn*. Available at: https://imbalanced-learn.org/stable/references/generated/imblearn.under_sampling.RandomUnderSampler.html. [Accessed 23 December 2024]
- [Jiang et al. (2021)] Jiang, W., Chen, J., Ding, X., Wu, J., He, J., & Wang, G.: “Review Summary Generation in Online Systems: Frameworks for Supervised and Unsupervised Scenarios”; *ACM Transactions on the Web*, 15, 3 (2021), 1–33.
- [Johar and Mubeen (2020)] Johar, S. and Mubeen, S.: “Sentiment Analysis on Large Scale Amazon Product Reviews”; *International Journal of Scientific Research in Computer Science and Engineering*, 8, 1 (2020), 7 - 15.
- [Kacem et al. (2020)] Kacem, A.B., Kamach, O., Chafik, S. and Hammou, M.A.: “A hybrid algorithm to size the hospital resources in the case of a massive influx of victims” *International Journal of Electrical and Computer Engineering (IJECE)*, 10, 1 (2020), 1006–1016.
- [Kausar, Fageeri and Soosaimanickam (2023)] Kausar, M.A., Fageeri, S.O. and Soosaimanickam, A.: “Sentiment Classification based on Machine Learning Approaches in Amazon Product Reviews”; *Engineering, Technology & Applied Science Research*, 13, 3 (2023), 10849–10855.
- [Li, Chow and Zhang (2020)] Li, N., Chow, C.-Y. and Zhang, J.-D.: “SEML: A Semi Supervised Multi-Task Learning Framework for Aspect-Based Sentiment Analysis”; *IEEE Access*, 8 (2020), 189287–189297.

- [Nellutla et al. (2021)] Nellutla, A. P., Hudnurkar, M., Ambekar, S. S., & Lidbe, A. D.: “Online Product Reviews and Their Impact on Third Party Sellers Using Natural Language Processing”; *International Journal of Business Intelligence Research*, 12, 1 (2021), 26–47.
- [Ngoc, Thu and Nguyen (2019)] Ngoc, T.N.T., Thu, H.N.T. and Nguyen, V.A.: “Mining aspects of customer’s review on the social network”; *Journal of Big Data*, 6, 1 (2019), 1–21.
- [Nkikabahizi, Cheruyiot and Kibe (2022)] Nkikabahizi, C., Cheruyiot, W. and Kibe, A.: “Chaining Zscore and feature scaling methods to improve neural networks for classification”; *Applied Soft Computing*, 123 (2022), 1–9.
- [Ni, Li and McAuley] Ni, J., Li, J., & McAuley, J.: “Justifying Recommendations using Distantly-Labeled Reviews and Fine-Grained Aspects”; “Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)”, (2019).
- [Park, Chung and Lee (2019)] Park, S., Chung, S., & Lee, S.: “The Effects of Online Product Reviews on Sales Performance: Focusing on Number, Extremity, and Length”; *Journal of Distribution Science*, 17, 5 (2019), 85–94.
- [Pennington, Socher and Manning (2014)] Pennington, J., Socher, R., & Manning, C. D.: “GloVe: Global Vectors for Word Representation”; (2014), 1–12.
- [Rachdian, Suryadi and Fransiscus (2022)] Rachdian, A.O., Suryadi, D. and Fransiscus, H.; “Identification of Customer Needs from Product Reviews using Topic Modeling and Aspect-based Sentiment Analysis”; *International Journal of Computing and Digital Systems*, 12, 6 (2022) 1383–1394.
- [Ramezani, Rahimi and Allan (2020)] Ramezani, S., Rahimi, R. and Allan, J.: “Aspect Category Detection in Product Reviews using Contextual Representation”; *Proceedings of ACM SIGIR Workshop on eCommerce, SIGIR eCom ’20 (2020)*, 1–6.
- [Sachin and Kumar (2022)] Sachin and Kumar, D.: “A Comprehensive Review on Data Classification Using Support Vector Machine”; *International Journal of Scientific Research in Engineering and Management (IJSREM)*, 6, 6 (2022), 1–10.
- [Scikit-Learn Developers (2024a)] Scikit-Learn Developers; “sklearn.ensemble.VotingClassifier”, Scikit Learn (2024). Available at: <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.VotingClassifier.html> [Accessed 23 December 2024].
- [Scikit-Learn Developers (2024b)] Scikit-Learn Developers; “sklearn.model_selection.RandomizedSearchCV”, Scikit-Learn (2024). Available at: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.RandomizedSearchCV.html [Accessed 23 December 2024].
- [Scikit-Learn Developers (2024c)] Scikit-Learn Developers; “sklearn.naive_bayes.GaussianNB”, Scikit Learn (2024). Available at: https://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.GaussianNB.html [Accessed 23 December 2024].
- [Scikit-Learn Developers (2024d)]. Scikit-Learn Developers; “StandardScaler”, Scikit-Learn (2024). Available at: <https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html> [Accessed 23 December 2024].
- [Sharma and Kaur (2021)] Sharma, T. and Kaur, K.: “Benchmarking Deep Learning Methods for Aspect Level Sentiment Classification”; *Applied Sciences*, 11, 22 (2021), 1–26.

- [Shetty, H and Mohammed (2025)] Shetty, A.M., H., M.D. and Mohammed, F.A.; “Fine-tuning XLNet for Amazon review sentiment analysis: A comparative evaluation of transformer models”; *ETRI Journal*, (2025), 1–18.
- [Shirahatti, Rajpurohit and Sannakki (2022)] Shirahatti, A., Rajpurohit, V. and Sannakki, S.; “Transformer based multi-head attention network for aspect-based sentiment classification”; *Indonesian Journal of Electrical Engineering and Computer Science*, 26, 1 (2022), 472–481.
- [Singh and Khandelwal (2024)] Singh, N.K. and Khandelwal, D.: “Sentiment analysis for amazon music instruments review.”; *International Journal of Research in Circuits, Devices and Systems*, 5, 1 (2024), 1–10.
- [Urkude et al. (2021)] Urkude, S., Hasanuzzaman, Urkude, V. and Kumar, C.: “Comparative Analysis on Machine Learning Techniques: A case study on Amazon Product Reviews”; *International Journal of Mechanical Engineering*, 6, 5 (2021), 739 - 744.
- [Xu and Ibrahim (2024)] Xu, Y. and Ibrahim, N.F.; “Cross-Domain Aspect-Based Sentiment Analysis for Enhancing Customer Experience in Electronic Commerce”; *Advances in Artificial Intelligence and Machine Learning Research*, 4, 3 (2024), 2593–2613.
- [Yang et al. (2019)] Yang, C., Zhang, H., Jiang, B. and Li, K.: “Aspect-based sentiment analysis with alternating coattention networks”; *Information Processing & Management*, 56, 3 (2019), 463–478.
- [Zhang, Li and Song (2019)] Zhang, C., Li, Q. and Song, D.: “Aspect-based Sentiment Classification with Aspect-specific Graph Convolutional Networks”; *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (2019), 4568–4578.
- [Zhang and Qiu (2021)] Zhang, G., & Qiu, H. “Competitive Product Identification and Sales Forecast Based on Consumer Reviews”; *Mathematical Problems in Engineering*, 2021 (2021), 1–15.
- [Zhu, Wang and Zou (2016)] Zhu, L., Wang, G. and Zou, X.: “A Study of Chinese Document Representation and Classification with Word2vec”; *Proceedings of the 2016 9th International Symposium on Computational Intelligence and Design (ISCID)*, (2016), 298–302.