# GenABSA-Vec: Generative Aspect-Based Sentiment Feature Vectorization for Document-Level Sentiment Classification

**Liu Minkang, Jasy Liew Suet Yan**
School of Computer Sciences, Universiti Sains Malaysia
11800 Penang, Malaysia
minkanglau21@student.usm.my, jasyliew@usm.my

## Abstract

Currently, document-level sentiment classification focuses on extracting text features directly using a deep neural network and representing the document through a high-dimensional vector. Such sentiment classifiers that directly accept text as input may not be able to capture more fine-grained sentiment representations based on different aspects in a review, which could be informative for document-level sentiment classification. We propose a method to construct a GenABSA feature vector containing five aspect-sentiment scores to represent each review document. We first generate an aspect-based sentiment analysis (ABSA) quadruple by finetuning the T5 pre-trained language model. The aspect term from each quadruple is then scored for sentiment using our sentiment lexicon fusion approach, SentLex-Fusion. For each document, we then aggregate the sentiment score belonging to the same aspect to derive the aspect-sentiment feature vector, which is subsequently used as input to train a document-level sentiment classifier. Based on a Yelp restaurant review corpus labeled with sentiment polarity containing 2040 documents, the sentiment classifier trained with ABSA features aggregated using geometric mean achieved the best performance compared to the baselines.

## 1 Introduction

Document-level sentiment analysis (DLSA) aims to detect the sentiment polarity of a document and is popularly used for product or service reviews (Liu, 2020). The outcomes of document-level sentiment classification could help individuals and businesses make more informed decisions based on user opinions and emotions (Onan, 2021; Zheng et al., 2020) especially in offline consumer consumption scenarios such as the selection of restaurants or entertainment venues. Sentiment polarity identified from reviews shows the general performance of a business, product or service, and provides useful information for consumers to uncover the opinions from previous customers (Bu et al., 2021; Le and Hui, 2022).

DLSA is framed as a conventional text binary classification task with the goal of identifying sentiment polarity (positive or negative) expressed in a unit of text. In the context of restaurant review, we denote one review text from a corpus as *t*, the word-based feature extraction method as *fea_ex*, and the different classification methods as *CLS*. Thus, the output is represented as $CLS(fea\_ex(t))$. Previous word-based feature representation methods focused on extracting sentiment at a coarse-grained level (i.e., directly from text) and may not capture necessarily relevant sentiment patterns or signals on finer-grained aspects causing the representation to be susceptible to spurious signals.

In contrast to DLSA, aspect-based sentiment analysis (ABSA) is a method that aims to analyze and understand user opinions at the aspect level (Zhang et al., 2023). ABSA enables the sentiment polarity detection of different objects on different attributes, thus allowing for fine-grained analysis within a document (Liu, 2020). ABSA can capture the sentiment score of each aspect in one review text. In general, ABSA contains multiple sub-tasks including Aspect Term Extraction (ATE), Aspect Category Detection (ACD), Opinion Term Extraction (OTE), and Aspect Sentiment Classification (ASC) (Zhang et al., 2023). The



*Proceedings of The 22nd Annual Workshop of the Australasian Language Technology Association*, pages 30–40

combined results from these four sub-tasks yield an ABSA quadruple to show a holder's specific opinion belonging to which aspect and towards which sentiment polarity.

When text is directly fed as the input into a document-level sentiment classification model, word-based features (i.e., lexical feature vectorization) represented in the form of a text embedding may not purely contain sentiment signals. Specifically, word-based features suffer from two main problems: 1) lack of explainability, and 2) high-dimensional feature space. The first problem is exacerbated with a wider adoption of neural embeddings capturing word relationships and semantics in numerical form automatically learnt from large corpora, thus resulting in the "black box" effect that makes text representations difficult to interpret (Arous et al., 2021; Zini and Awad, 2022). The second problem is correlated with the growing amount of text data used in sentiment classifiers. As the size of a corpus increases, the dimensionality of text data also increases exponentially, which can lead to the curse of dimensionality and make it difficult for certain machine learning models to reach convergence during training (Chang et al., 2020). DLSA and ABSA have usually been addressed as two separate tasks in the realm of sentiment analysis and have never been fused before.

Our main goal in this paper is to test if using ABSA-generated (GenABSA) features to represent a review can more succinctly capture important sentiment signals from text to improve DLSA. Each review is represented by a fixed-length vector, containing only the sentiment score on five selected aspects in the restaurant domain. We ran experiments on the Yelp restaurant domain corpus to classify the sentiment polarity (positive or negative) of a review given a GenABSA feature vector, $<score_{food}$, $score_{service}$, $score_{ambience}$, $score_{location}$, $score_{drink}>$, computed using different feature aggregation methods.

## 2 Related Work

### 2.1 Document-level Sentiment Analysis (DLSA)

For DLSA on user-generated review text, the text representation mostly comes from direct encoding of the review text although there has been attempts to integrate with user and product embeddings (Lyu et al., 2020). Prior DLSA studies focused on fusing different network frameworks or machine learning methods to extract more accurate features to be fed to the sentiment classifier. Tripathy et al. (2017) applied a two-step hybrid approach to detect the sentiment polarity of each document. Support vector machine was first used to select important features from a document, and then the selected features were sent to a neural network for sentiment classification. Rao et al. (2018) proposed a long short-term memory (LSTM) framework with two hidden layers to extract sentiment polarity. The first hidden layer represented each sentence, and the second layer encoded the document representation.

Blended deep learning frameworks have also been employed to address DLSA. Rhanoui et al. (2019) proposed a CNN-BiLSTM model for sentiment classification. The CNN convolution layer was used to extract a maximum amount of information from the document while the BiLSTM layer processed the output from the convolution layer from a time-series perspective. Subsequently, the classification result was obtained through a softmax output layer. Due to the poor adaptability of the existing sentiment lexicons, Sun et al. (2019) constructed a model combined with domain-specific sentiment words for DLSA, which classified each document based on a combination of document and emotion features. Document features were generated by Asymmetric Convolutional Neural Network (ACNN) and word and sentence features were extracted using Bidirectional Gated Recurrent Neural Network (BGRNN). Emotion features were generated by a domain-specific sentiment lexicon. Onan (2021) used TF-IDF weighted GloVe word embedding combined with 1-3 grams convolution to extract features from a document, and a LSTM layer to encode the features. The model fused more deep-learning components to obtain a representation of the document.

Liu et al. (2020) proposed the AttDR-2DCNN model to take advantage of the attention mechanism in identifying important words and sentences for sentiment classification, followed by a two-dimensional convolution layer and Convolution Block Attention Module (CBAM) to further extract features. On the other hand, Zhang et al. (2021b) employed attention mechanism with BiLSTM to select the most critical tokens in the

documents, and gradually downsized the scale of the document to overcome the problem of the model paying more attention to the tail words.

In contrast, Atandoh et al. (2023) integrated a pre-trained BERT with a one-gram convolution neural network layer for sentiment classification. BERT was used for encoding the words in the document while the CNN layer was responsible to further extract key features for sentiment analysis. Compared with the conventional embedding methods, BERT pre-trained on a large amount of text can obtain more accurate results in the downstream sentiment classification task. Although most prior studies incorporated feature extraction within a neural network architecture, Wasi and Abulaish (2024) performed feature extraction by injecting general knowledge and domain-specific knowledge to generate fusion features for a logistic regression model.

### 2.2 Aspect-based Sentiment Analysis (ABSA)

ABSA is typically framed as a triplet extraction or quadruple extraction task. A triplet consists of an aspect category, aspect term, and sentiment polarity of the aspect term. In contrast, a quadruple has an additional element (i.e., opinion term).

Joint element detection focuses on target and sentiment polarity detection for the ABSA task but still does not concurrently produce all elements of the triplet or quadruple. Therefore, ABSA can be framed as a multi-task framework to obtain all the elements of the triplet or quadruple at the same time. He et al. (2019) proposed an interactive multi-task learning network (IMN) including aspect term and opinion term co-extraction, aspect-level sentiment classification, document-level sentiment, and document-level domain classification. The framework accepted a sequence as input and took advantage of message-passing graphical model inference algorithms to allow informative interactions between sub-tasks. Zhao et al. (2023) proposed a multitask learning model combining aspect polarity classification (APC) and aspect term extraction (ATE) sub-tasks. These two sub-tasks encoded the tokens using BERT. ATE was obtained using a linear layer. For APC, the framework added a multi-head attention (MHA) module to enhance the connection between aspects and their associated dependencies to obtain a more informative representation for classification. Some methods directly used BERT as the main component to obtain the ABSA results. Li et al. (2019) exploited BERT as the embedding layer to represent the text input, which was connected to different layers to obtain the ABSA results. Such method eliminates the need to design a complicated network to match the ABSA sub-tasks.

As BERT is pre-trained with text from general domains, it may not generalize well on product reviews from specific domains such as restaurant, hotel, and electronic product. DomBERT was designed to address this problem by first classifying text as belonging to which domain and then extended the BERT on in-domain corpus and relevant domain corpora before being used with a classifier layer to generate the ABSA results (Xu et al., 2020).

With the rapid development of large language models (LLMs), ABSA has also recently been formulated using a generative approach. A generative model for ABSA takes in the original text as input to concurrently generate a triplet or quadruple containing the desired sub-tasks. For quadruple extraction using the generative paradigm, Zhang et al. (2021a) employed paraphrase generation to define the ABSA output format in natural language. The model for quadruple generation was obtained by finetuning the parameters of a T5 pre-trained language model. In this paper, we explored the generative approach to produce ABSA quadruples.

## 3 Methodology

Figure 1 shows our methodological framework encompassing four phases: 1) ABSA quadruple generation, 2) ABSA feature extraction, and 3) ABSA feature aggregation and 4) document-level sentiment classification.

We first use a generative method to extract ABSA quadruples from each review. To obtain aspect-based sentiment features, the opinion term from each ABSA quadruple is then scored for sentiment. It is possible for multiple ABSA quadruples to be generated for a single review. Therefore, the sentiment scores from all quadruples in each review are aggregated based on five aspects of interest (i.e., food, service, ambience, location, and drink) into a feature

vector containing five elements to serve as input to a document-level sentiment classifier.

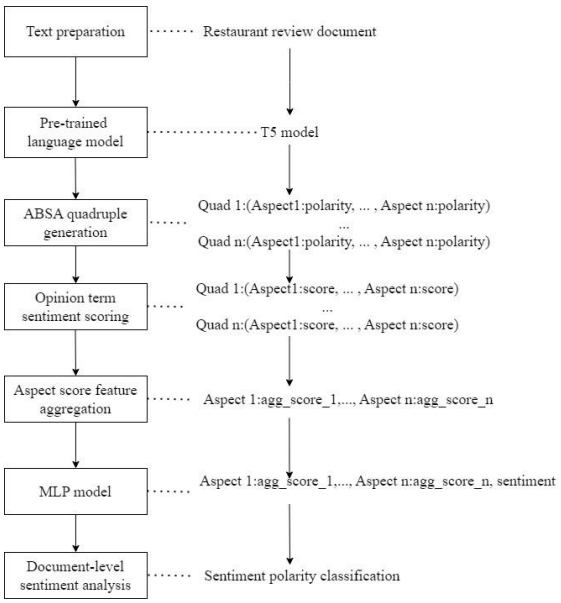

Figure 1: Methodological framework.

### 3.1 Dataset

For ABSA, we used the SemEval-2016 Task 5 restaurant domain dataset (Pontiki et al., 2016) to finetune our ABSA quadruple generation model. The SemEval-2016 Task 5 dataset contains annotations on the aspect category, opinion target expression and sentiment polarity[1].

For DLSA, we created a new dataset for document-level sentiment analysis by randomly selecting restaurant reviews from the Yelp Open Dataset[2]. Each review includes a user id, business id, review date, review text, and a score rated by a user in a range from 1 to 5. We followed the same method as Blitzer et al. (2007) to label the reviews with user rating score > 3 as positive and reviews with the score < 3 as negative. The rest of the reviews were discarded. Our final Yelp dataset contains 1020 positive samples and 1020 negative samples (i.e., balanced class distribution). The generation of the ABSA quadruples from each review document is resource- and time-intensive so we chose a reasonably sized test set to make systematically running experiments with various configurations feasible.

### 3.2 ABSA Quadruple Generation

We employed a generative method to obtain the ABSA quadruples in each document by processing every sentence. The generative method formulates the ABSA task as a text-to-text method and finetunes a T5 pre-trained language model (i.e., T5-base[3]). For the ABSA generative model, we made the original text review as the input of the T5 pre-trained model, and the quadruple containing aspect category, aspect term, opinion term, and aspect sentiment polarity as the output. For example, given an input text "serves really good sushi", the output is {"aspect": "sushi", "opinion": "good", "polarity": "positive", "category": "FOOD"} ({aspect term, opinion term, sentiment polarity, aspect category}).

For ABSA quadruple generation, we used 1530 samples from the SemEval-2016 Task 5 dataset as the finetuning set and 583 samples as the test set. The optimal ABSA model hyperparameters are shown in Table 1.

| Hyperparameter | Value |
|---|---|
| Learning rate | 5e-5 |
| Batch size | 10 |
| Epoch | 30 |
| Weight decay | 0.01 |

Table 1: Optimal ABSA model hyperparameters.

Based on the test set, the ABSA quadruple extraction model achieved an accuracy of 0.74 and a macro F1 score of 0.52 as shown in Table 2. The model performance is computed across all four sub-tasks. We then applied the finetuned ABSA model to generate the ABSA quadruples for each restaurant review in the Yelp dataset.

| Metric | Score |
|---|---|
| Accuracy | 0.7419 |
| Macro-Precision | 0.5210 |
| Macro-Recall | 0.5267 |
| Macro-F1 | 0.5214 |

Table 2: Evaluation metrics on the ABSA quadruple extraction task.

[1] https://alt.qcri.org/semeval2016/task5/

[2] https://www.yelp.com/dataset

[3] https://huggingface.co/google-t5/t5-base

### 3.3 ABSA Feature Extraction

The generated ABSA quadruples or quads are used to construct aspect-sentiment document features for sentiment analysis. We follow the aspect categories used in the restaurant domain of SemEval-2016 Task 5 (Pontiki et al., 2016), which consists of five aspects including food, services, ambience, location, and drinks. Therefore, each review document is represented using these five aspects with each aspect being assigned a corresponding sentiment valence. As the quad returns only the opinion term and sentiment polarity, we derive a sentiment score based on the opinion term by referencing the valence of the opinion word from a fusion of sentiment lexicons, SentLex-Fusion. SentLex-Fusion is a fusion of four sentiment lexicons shown in Table 3 to maximize the coverage of opinion terms to be scored for sentiment.

| Lexicon | Size | Description |
|---|---|---|
| AFINN[4] (Nielsen, 2011) | 3382 | S: [-5, 5]<br>V: 1.65 |
| SO-CAL[5] (Taboada et al., 2011) | 6395 | S: [-5, 5]<br>V: 1.11 |
| WKWSCI[6] (Khoo and Johnkhan, 2018) | 29914 | S: [-3, 3] [it range], [-2, 2] [ph range] |
| SentiWordNet[7] (Baccianella et al., 2010) | 117660 | S: [-5, 5]<br>V: 3.0 |
| SentLex-Fusion | 100170 | S: [-5, 5] |

Table 3: Description of sentiment lexicons in SentLex-Fusion (S = Polarity score range, V = Version, it = individual term, ph = phrase).

In the fusion stage, we integrate all the terms from all four lexicons, filter duplicate terms, and map different score ranges into a standardized range of [-5, 5]. If an opinion term occurs in more than one lexicon, the average sentiment score for the opinion term is calculated as the final score in SentLex-Fusion. After fusion, SentLex-Fusion contains 100170 terms, which is five times the coverage of opinion terms found in the ABSA quads generated from the Yelp dataset (17675 opinion terms). The coverage percentage of in-lexicon opinion terms is 90.6%.

However, we discovered two problems in the process of scoring the opinion terms. First, not all the opinion terms in the ABSA quads are within the coverage of SentLex-Fusion. Of the 17675 opinion terms, we found 1670 out-of-lexicon (OOL) terms without corresponding terms in SenLex-Fusion, indicating 9.4% opinion terms require sentiment imputation. We illustrate the problem using Example 1.

Example 1 (Sentence): *The fish was truly ambrosial, while the beer was delightful.*

Two quadruples are extracted from Text 1:
Quad 1: [‘aspect’: fish, ’polarity’: positive, ‘opinion’: **ambrosial**, ‘category’: Food]
Quad 2: [‘aspect’: *beer*, ’polarity’: positive, ‘opinion’: **delightful,** ‘category’: Drink]

Using SentLex-Fusion, the opinion term “delightful” can be mapped to a sentiment score of 3.67, but no matching word from the lexicon can be found for the opinion term “ambrosial”. To overcome the first problem, we designed an imputation method to handle opinion words that cannot be matched to SentLex-Fusion. For opinion terms not found in SentLex-Fusion, we generated a score heuristically based on sentiment polarity. If the sentiment polarity is positive, we assign +3 as the score to replace the opinion term. If the sentiment polarity is negative, the score is set to -3. Based on the SentLex-Fusion valence scale, 3 is the midpoint value of the positive scale range, and -3 corresponds to the midpoint of the negative scale range. In addition, a sentiment score of 0 is assigned to aspect categories that are absent from a review. In Example 1, after applying our imputation rules, the review text is represented as an ABSA feature vector of [3, 0, 0, 0, 3.67] ([food, service, ambience, location, drink]).

The second problem is caused by the value 0 in the ABSA feature vector holding two possible meanings. An aggregated aspect-sentiment score of 0 could mean the absence of an aspect category

[4] https://github.com/fnielsen/afinn/blob/master/afinn/data/AFINN-en-165.txt
[5] https://github.com/sfu-discourse-lab/SOCAL/tree/master/Resources/dictionaries
[6] https://researchdata.ntu.edu.sg/dataset.xhtml?persistentId=doi:10.21979/N9/DWWEBV
[7] https://wordnet.princeton.edu/download/current-version

in a review or it could also mean two or more opinion terms with opposite polarity of the same aspect category within a review summing up to 0. Example 2 illustrates the problem.

For service and drink, SentLex-Fusion can assign sentiment scores to these two aspect categories. However, for food, both "epicurean" and "woefully insipid" have no match found in SentLex-Fusion. Using our imputation rules, the opinion term "epicurean" corresponds to +3, while "woefully insipid" corresponds to -3. In the aggregation stage, if one aspect category includes multiple scores, the mean aspect-sentiment score is computed. As a result, Example 2 is represented by an ABSA feature vector of [0, 2.84, 0, 0, 2.25]. Although the food aspect category occurs as captured by the two ABSA quadruples in Example 2, the final aggregated food aspect sentiment score of 0 implies the absence of the food aspect in the review, thus causing inaccuracies in the ABSA feature representation.

Example 2 (Sentence): *The steak was an epicurean, while the chicken was woefully insipid, but the staff is nice and the juice is great.*

Four quadruples are extracted from Text 2:
Quad 1: ['aspect': steak, 'polarity': positive, 'opinion': **epicurean**, 'category': Food]
Quad 2: ['aspect': chicken, 'polarity': negative, 'opinion': **woefully insipid**, 'category': Food]
Quad 3: ['aspect': staff, 'polarity': positive, 'opinion': **nice**, 'category': Service]
Quad 4: ['aspect': juice, 'polarity': positive, 'opinion': **great**, 'category': Drink]

To avoid ambiguity caused by the double meaning of 0, we applied feature scaling to adjust the original scale range to a positive range. We maintained the actual range of the original scale but shifted to a positive scale (i.e., -5 is mapped to 1 and 5 to 11 with 6 now being the midpoint replacing the original 0 so 6 represents neutral and 0 now carries no sentiment).

Equation 1 is used to adjust the scale of the sentiment score from a value between -5 to 5 to the range from 1 to 11.

$$X_{new} = \frac{(X - from\_min) \times (to\_max - to\_min)}{(from\_max - from\_min)} + to\_min \quad (1)$$

In Example 2, the new ABSA feature vector is computed as [6, 8.84, 0, 0, 8.25] (from_min = -5, from_max = 5, to_max = 11, and to_min = 1). The food aspect sentiment is assigned to a neural score of 6 instead of 0.

### 3.4 ABSA Feature Aggregation

As one review document may contain more than one ABSA quad, we introduced two aggregation methods in our experiments.

**Method 1: Simple Mean**

The first aggregation method simply applies a simple mean to the sum of each aspect's sentiment scores within a document. Suppose $score_i^{food}$, $score_i^{service}$, $score_i^{ambience}$, $score_i^{location}$, and $score_i^{drink}$ denote the aspect sentiment score of food, service, ambience, location, and drink. The sum of sentiment scores for each aspect is divided by *n* number of times an aspect category is mentioned in a review. The simple average to compute the aggregated ABSA feature vector is illustrated in Equation 2.

$$\left[\frac{\sum_{i=1}^{n} score_i^{food}}{n^{food}}, \frac{\sum_{i=1}^{n} score_i^{service}}{n^{service}}, \frac{\sum_{i=1}^{n} score_i^{ambience}}{n^{ambience}}, \frac{\sum_{i=1}^{n} score_i^{location}}{n^{location}}, \frac{\sum_{i=1}^{n} score_i^{drink}}{n^{drink}}\right] \quad (2)$$

**Method 2: Geometric Mean**

Geometric mean captures serial correlation in a variable. Specifically, geometric mean measures the relationship between a variable's current value given its past values (Ando et al., 2004). The occurrence of one aspect multiple times in a review may be correlated and this feature aggregation method can capture that correlation.

For the aspect-sentiment score generated by SentLex-Fusion, the geometric mean can maintain negative values representing negative sentiment and positive values representing positive sentiment as the original scale [-5,5] is adjusted to [-1,1]. Additionally, a plus point of geometric mean is it does not assign 0 to the aspect mentioned in the review text to avoid ambiguity.

The ABSA feature aggregation method using geometric mean follows two steps. In Step 1, we apply absolute maximum scaling as shown in

Equation 3 to convert the scale of -5 to 5 into -1 to 1. We need to find the absolute maximum value of the feature in the dataset and divide all the values in the column by that maximum value.

$$X_{new} = \frac{X}{|max(X)|} \quad (3)$$

In Step 2, we apply geometric mean on the sentiment scores for each aspect category. Equation 4 shows the geometric mean calculation for one aspect category (i.e., food aspect). We add 1 to each sentiment score to avoid any problems with negative percentages but subsequently subtract 1 from the result. To illustrate feature aggregation using geometric mean, suppose we extract four quads from a review text, and the four sentiment scores are related to the food aspect, Equation 4 computes the aggregated food aspect sentiment score using geometric mean. The computation for other aspect categories follows the same equation.

$$Geometric\ mean_{food} = \left[\left(1 + score_1^{food}\right) \times \left(1 + score_2^{food}\right) \times \left(1 + score_3^{food}\right) \times \left(1 + score_4^{food}\right)\right]^{\frac{1}{4}} - 1 \quad (4)$$

### 3.5 Document-Level Sentiment Classifier

The extracted ABSA feature vector serves as input to the document-level binary sentiment classifier. For DLSA, we utilize a multi-layer perceptron (MLP) to classify the sentiment polarity of each document.

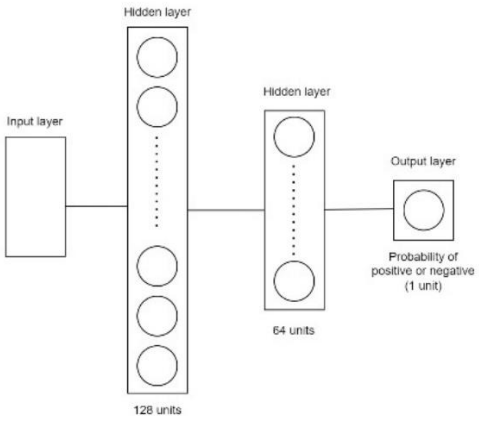


Figure 2: MLP architecture for document-level sentiment classification.

The MLP architecture comprises one input layer with 128 dimensions, two hidden layers (the first hidden layer has 128 neurons and the second hidden layer has 64 neurons) with ReLU activation function, and one output layer with sigmoid activation function. We set the training epoch to 20, the dropout probability to 0.3 to prevent overfitting, and the loss function to binary cross-entropy.

### 3.6 DLSA Experiment Setup

We propose one feature scaling method and two sentiment score aggregation methods (i.e., simple mean and geometric mean), thus producing three different GenABSA feature vectors to be examined in our experiments: 1) SentLex-Fusion ABSA features without feature scaling aggregated using simple mean (**ABSA + SM**), 2) SentLex-Fusion ABSA features with feature scaling aggregated using simple mean (**ABSA + FS + SM**), and 3) SentLex-Fusion ABSA features using geometric mean (**ABSA + GM**). The three GenABSA feature vectors are evaluated using the same MLP model for document-level sentiment analysis. We split the 2040 samples (1020 positive and 1020 negative) into a training set, validation set, and test set following the ratio of 8:1:1. We have selected accuracy as our primary performance metric as the binary sentiment classes are evenly distributed.

Our GenABSA feature-based sentiment classifiers are then compared with the four following baselines commonly found in studies on document-level sentiment analysis of reviews (Rao et al., 2018; Atandoh et al., 2023; Tripathy et al., 2017). The baseline models directly extract text features from the reviews.

[1] TF-IDF with MLP (**TF-IDF + MLP**): We utilize TF-IDF to represent the features of a review document, which is then connected to a MLP for sentiment classification. The input dimension is set to 800.

[2] Word2Vec with MLP (**W2V + MLP**): A custom Word2Vec word embedding is first trained the Yelp dataset and then used to extract the document representations to be fed into the MLP. The input dimension is set to 100.

[3] BERT with MLP (**BERT + MLP**): BERT[8] is used to extract the document representations for

[8] https://huggingface.co/google-bert/bert-base-uncased

the MLP sentiment classifier. The input dimension is set to 768.

[4] Finetuning a pre-trained sentiment analysis model (**PRE-SENT**): This method directly finetunes an existing pre-trained sentiment analysis model[9]. No connection is needed to MLP. Instead, we perform finetuning directly on the pre-trained sentiment analysis model to update the model parameters.

## 4 Results and Discussion

Table 4 shows the document-level sentiment classification performance comparison between our GenABSA feature vectors and the four baselines. The scores in bold represent the best-performing model.

| Method | A | P | R | F1 |
|---|---|---|---|---|
| ABSA + SM | 0.915 | 0.912 | 0.913 | 0.913 |
| ABSA + FS + SM | 0.909 | 0.904 | 0.915 | 0.907 |
| ABSA + GM | **0.941** | **0.943** | **0.936** | **0.939** |
| TF-IDF + MLP | 0.915 | 0.915 | 0.915 | 0.915 |
| W2V + MLP | 0.789 | 0.791 | 0.789 | 0.789 |
| BERT + MLP | 0.913 | 0.914 | 0.913 | 0.913 |
| PRE-SENT | 0.917 | 0.924 | 0.917 | 0.916 |

Table 4: Document-level sentiment model performance (A = Accuracy, P = Macro-Precision, R = Macro-Recall, F1 = Macro-F1).

Of the three GenABSA feature vectors, ABSA + GM produced the best results, which proves that using geometric mean as the ABSA feature aggregation method is more effective than merely using simple mean. Surprisingly, our feature scaling method to differentiate between neutral sentiment and no sentiment leads to a slight decrease in model accuracy, precision, and F1. This could mean capturing neutral sentiment in the ABSA vectors counterintuitively added a layer of complexity and confusion to the sentiment classification model.

Based on the evaluation metrics, the GenABSA feature-based models yield comparable performance to the baselines. ABSA + GM achieved the highest accuracy, precision, recall and F1, outperforming all the baselines. Our GenABSA models successfully achieved competitive performance to baselines with a low-dimensional feature vector containing only five dimensions as opposed to higher-dimensional text vectors. ABSA + GM not only achieved notable improvements over the simple and naïve text representations such as TF-IDF and Word2Vec but also outperform the richer text representations such as BERT and the pre-trained sentiment analysis model which presumably have been pre-trained with larger external resources for the sentiment classification task. This finding implies that aspect-sentiment features can semantically capture more meaningful sentiment signals with reduced noise, thus increasing the likelihood for the sentiment classifier to learn more succinct patterns to distinguish between the two sentiment classes.

In fact, the aspect-sentiment features are more explainable as illustrated by Example 3 and Example 4 compared to the more complex textual embeddings. It is easy to explain ABSA + GM classified the review in Example 3 as positive because of the positive sentiment scores for food, service and ambience whereas the negative sentiment scores for these three aspects led to the review in Example 4 being classified as negative.

Example 3 (Review): *Best Thai food in Santa Barbara area. Well priced, great outdoor area. Casual and easy. Takeout is always on point. What more could you want from a mid-priced Thai restaurant in a small beach community?*

ABSA + GM Feature Vector:
[0.733, 0.600, 0.233, 0, 0]
([food, service, ambience, location, drink])
Actual: Positive; Predicted: Positive

Example 4 (Review): *Horrible customer service at this Logan's location. I've had mixed experiences with each visit but this was by far the worst. Against better judgement, I returned after*

[9] https://huggingface.co/prasadsawant7/sentiment-analysis-pretrained/tree/main

*being served burnt food and waiter argued the food was not burnt. Poor quality, poor customer service and filthy bathrooms. (Failed to mention, bathroom horribly dirty, broken blocks on doors and broken toilet seats. Reminds me of the bathroom in a public park overrun with the homeless).*

ABSA + GM Feature Vector:
[-0.262, -0.409, -0.455, 0, 0]
([food, service, ambience, location, drink])
Actual: Negative; Predicted: Negative

Despite GenABSA's strength in terms of explainability, our preliminary error analysis on misclassified examples reveals its sensitivity towards explicit sentiment terms tied to a specific aspect in a sentence as illustrated in Example 5. GenABSA focused only on the phrase *"suck good cookies"*, which produced a positive food sentiment score albeit being low but missed other sentiment signals (e.g., *"such sneaks"*, *"very disheartening"*) from short sentences without reference to any specific aspect.

Example 5 (Review): *I still have not learned my lesson. I stopped in there to buy cookies for my son because they always had suck good cookies. I bought a box of them off the counter. Well, I get home, open them and they are steal! and there were xmas cookies hidden under the other cookies. Yes, xmas cookies hidden in bottom of box. Such sneaks! So sad what this place has become. Seriously? Its the middle of february, past the middle. Very disheartening!!!!!!!!!!!!!!!!!!!*

ABSA + GM Feature Vector:
[0.167, 0, 0, 0, 0]
([food, service, ambience, location, drink])
Actual: Negative; Predicted: Positive

## 5 Conclusion and Future Work

In conclusion, instead of following the typical text feature extraction pipeline for DLSA, we experimented with a more novel GenABSA approach to first extract ABSA features using a generative model and then aggregating the aspect-sentiment signals into a more compact ABSA feature vector for the downstream document-level sentiment classification task. Our main contribution in this paper is to provide empirical insights on how to extract aspect-sentiment information from generated ABSA quadruples to be transformed into a compact ABSA feature vector that would serve as the most effective aspect-sentiment feature representation for DLSA. Our findings show our low-dimensional ABSA feature vectors yield at par performance with baselines using text features. We also found that geometric mean has demonstrated more promising results compared to using simple mean in ABSA feature aggregation.

Our study has proven it is possible to fuse ABSA (i.e., extracting aspect-sentiment signals first from text) into the DLSA pipeline with promising results. We have yet to thoroughly examine the feasibility of the method in terms of computational time on a large dataset as opposed to using direct text input and conduct an error analysis based on the performance of each ABSA sub-task, which leaves room for our future work. Future research efforts can also investigate the application and finetuning of other LLMs for ABSA quadruple generation to capture aspect-sentiment signals more accurately. In addition, other imputation methods can be explored to fill the missing sentiment scores caused by the coverage of the lexicon.

## 6 Limitations

First, we only focused on the restaurant domain in this study. The restaurant domain uses a limited set of aspect categories that may be hard to adapt to other domains. As such, our findings in the paper may not generalize to other domains as the methodology has yet to be tested on other domains. Second, we limited sentiment polarity in the DLSA task to only positive and negative, so further exploration is required to apply the methodology to scenarios that include neutral sentiment and no sentiment. Third, we limited the size of our Yelp restaurant review test set for the DLSA evaluation to make running extensive experiments feasible, which might have limited the generalizability of our GenABSA models on a larger variety of restaurant reviews.

## Acknowledgments

This study was funded by "Ministry of Higher Education Malaysia for Fundamental Research Grant Scheme with Project Code: FRGS/1/2020/ICT02/USM/02/3".

## References


T. Ando, Chi-Kwong Li, and Roy Mathias. 2004. Geometric means. *Linear algebra and its applications*, 385:305–334.

Ines Arous, Ljiljana Dolamic, Jie Yang, Akansha Bhardwaj, Giuseppe Cuccu, and Philippe Cudré-Mauroux. 2021. MARTA: Leveraging human rationales for explainable text classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(7):5868–5876.

Peter Atandoh, Fengli Zhang, Daniel Adu-Gyamfi, Paul H. Atandoh, and Raphael Elimeli Nuhoho. 2023. Integrated deep learning paradigm for document-based sentiment analysis. *Journal of King Saud University - Computer and Information Sciences*, 35(7):101578.

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining.

John Blitzer, Mark Dredze, and Fernando Pereira. 2007. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In Annie Zaenen and Antal van den Bosch, editors, *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 440–447, Prague, Czech Republic. Association for Computational Linguistics.

Jiahao Bu, Lei Ren, Shuang Zheng, Yang Yang, Jingang Wang, Fuzheng Zhang, and Wei Wu. 2021. ASAP: A Chinese review dataset towards aspect category sentiment analysis and rating prediction. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2069–2079, Online. Association for Computational Linguistics.

Jing-Rong Chang, Hsin-Ying Liang, Long-Sheng Chen, and Chia-Wei Chang. 2020. Novel feature selection approaches for improving the performance of sentiment classification. *Journal of Ambient Intelligence and Humanized Computing*.

Ruidan He, Wee Sun Lee, Hwee Tou Ng, and Daniel Dahlmeier. 2019. An interactive multi-task learning network for end-to-end aspect-based sentiment analysis. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 504–515, Florence, Italy. Association for Computational Linguistics.

Christopher Sg Khoo and Sathik Basha Johnkhan. 2018. Lexicon-based sentiment analysis: Comparative evaluation of six sentiment lexicons. *Journal of Information Science*, 44(4):491–511.

Tan Khang Le and Siu Cheung Hui. 2022. Machine learning for food review and recommendation. arXiv:2201.10978 [cs].

Xin Li, Lidong Bing, Wenxuan Zhang, and Wai Lam. 2019. Exploiting BERT for end-to-end aspect-based sentiment analysis. In Wei Xu, Alan Ritter, Tim Baldwin, and Afshin Rahimi, editors, *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 34–41, Hong Kong, China. Association for Computational Linguistics.

Bing Liu. 2020. *Sentiment analysis: Mining opinions, sentiments, and emotions*. Cambridge University Press, Second Edition.

Fagui Liu, Jingzhong Zheng, Lailei Zheng, and Cheng Chen. 2020. Combining attention-based bidirectional gated recurrent neural network and two-dimensional convolutional neural network for document-level sentiment classification. *Neurocomputing*, 371:39–50.

Chenyang Lyu, Jennifer Foster, and Yvette Graham. 2020. Improving document-level sentiment analysis with user and product context. In Donia Scott, Nuria Bel, and Chengqing Zong, editors, *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6724–6729, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Finn Årup Nielsen. 2011. A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. In *Proceedings of the ESWC2011 Workshop on Making Sense of Microposts*, pages 93–98, Heraklion, Crete, Greece.

Aytuğ Onan. 2021. Sentiment analysis on product reviews based on weighted word embeddings and deep neural networks. *Concurrency and Computation: Practice and Experience*, 33(23):e5909.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad AL-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, Véronique Hoste, Marianna Apidianaki, Xavier Tannier, Natalia Loukachevitch, Evgeniy Kotelnikov, Nuria Bel, Salud María Jiménez-Zafra, and Gülşen Eryiğit. 2016. SemEval-2016 task 5: Aspect based sentiment analysis. In Steven Bethard, Marine Carpuat, Daniel Cer, David Jurgens, Preslav Nakov, and Torsten Zesch, editors, *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 19–30, San Diego, California. Association for Computational Linguistics.

Guozheng Rao, Weihang Huang, Zhiyong Feng, and Qiong Cong. 2018. LSTM with sentence representations for document-level sentiment classification. *Neurocomputing*, 308:49–57.

Maryem Rhanoui, Mounia Mikram, Siham Yousfi, and Soukaina Barzali. 2019. A CNN-BiLSTM model for document-level sentiment analysis. *Machine Learning and Knowledge Extraction*, 1(3):832–847.

Chengai Sun, Fang Wang, and Gang Tian. 2019. Document-level sentiment analysis based on domain-specific sentiment words. *Journal of Physics: Conference Series*, 1288(1):012052.

Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 37(2):267–307.

Abinash Tripathy, Abhishek Anand, and Santanu Kumar Rath. 2017. Document-level sentiment classification using hybrid machine learning approach. *Knowledge and Information Systems*, 53(3):805–831.

Nesar Ahmad Wasi and Muhammad Abulaish. 2024. SKEDS — An external knowledge supported logistic regression approach for document-level sentiment classification. *Expert Systems with Applications*, 238:121987.

Hu Xu, Bing Liu, Lei Shu, and Philip Yu. 2020. DomBERT: Domain-oriented language model for aspect-based sentiment analysis. In Trevor Cohn, Yulan He, and Yang Liu, editors, *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1725–1731, Online. Association for Computational Linguistics.

Wenxuan Zhang, Yang Deng, Xin Li, Yifei Yuan, Lidong Bing, and Wai Lam. 2021a. Aspect sentiment quad prediction as paraphrase generation. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9209–9219, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Wenxuan Zhang, Xin Li, Yang Deng, Lidong Bing, and Wai Lam. 2023. A survey on aspect-based sentiment analysis: Tasks, methods, and challenges. *IEEE Transactions on Knowledge and Data Engineering*, 35(11):11019–11038.

You Zhang, Jin Wang, and Xuejie Zhang. 2021b. Conciseness is better: Recurrent attention LSTM model for document-level sentiment analysis. *Neurocomputing*, 462:101–112.

Guoshuai Zhao, Yiling Luo, Qiang Chen, and Xueming Qian. 2023. Aspect-based sentiment analysis via multitask learning for online reviews. *Knowledge-Based Systems*, 264:110326.

Lin Zheng, Naicheng Guo, Weihao Chen, Jin Yu, and Dazhi Jiang. 2020. Sentiment-guided sequential recommendation. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1957–1960, New York, NY, USA. Association for Computing Machinery.

Julia El Zini and Mariette Awad. 2022. On the explainability of natural language processing deep models. *ACM Computing Surveys*, 55(5):103:1-103:31.