Interdisciplinary Journal of Information, Knowledge, and Management

An Official Publication of the Informing Science Institute InformingScience.org

IJIKM.org

Volume 20, 2025

BERT-BASED MODEL AND LLMS-GENERATED SYNTHETIC DATA FOR CONFLICT SENTIMENT IDENTIFICATION IN ASPECT-BASED SENTIMENT ANALYSIS

Nuryani*	School of Electrical and Informatics Engineering, Bandung Institute of Technology (ITB), Bandung, Indonesia, and National Research and Innovation Agency (BRIN), Indonesia	33219005@std.stei.itb.ac.id
Rinaldi Munir	School of Electrical and Informatics Engineering, Bandung Institute of Technology (ITB), Bandung, Indonesia	rinaldi@informatika.org
Ayu Purwarianti	School of Electrical and Informatics Engineering, Bandung Institute of Technology (ITB), Bandung, Indonesia	ayu@itb.ac.id
Dessi Puji Lestari	School of Electrical and Informatics Engineering, Bandung Institute of Technology (ITB), Bandung, Indonesia	dessipuji@informatika.org

* Corresponding author

ABSTRACT

Aim/Purpose

Most research in sentiment analysis, as well as aspect-based sentiment analysis (ABSA), classifies sentiment polarity into two classes (positive and negative) or three classes (positive, negative, and neutral), excluding conflict sentiment. A sentiment will be classified as conflict if it expresses both positive and negative sentiments. Ignoring conflict sentiment will cause the classification to be less accurate. This study investigates the four-class sentiment classification (positive, negative, neutral, and including conflict) and proposes a model utilizing a pre-trained language representation model (BERT) for identifying conflict sentiment in ABSA. We also employ an open-source large language model (LLM)

Accepting Editor Narongsak Sukma | Received: October 1, 2024 | Revised: December 19, 2024; January 9, January 13, 2025 | Accepted: January 13, 2025.
Cite as: Nuryani, Munir, R., Purwarianti, A., & Lestari, D. P. (2025). BERT-based model and LLMs-generated synthetic data for conflict sentiment identification in aspect-based sentiment analysis. *Interdisciplinary Journal of Information, Knowledge, and Management, 20,* Article 4. https://doi.org/10.28945/5439

	created by Meta, Llama 3, for generating synthetic data to support research on four-class sentiment classification in ABSA.
Background	Public opinions and experiences on product reviews, social events, political movements, etc., can be used for exploring customer behavior, predicting customer preferences, understanding public sentiment, etc., so it becomes an important component in the decision-making process. Providing an accurate opinion will enable an individual, business, or organization to have an informed judgement before making a decision. An aspect-based sentiment analysis, utilizing a four-class sentiment classification system - comprising positive, negative, neutral, and conflict - will produce a more precise assessment than a general sentiment analysis utilizing a two- or three-class sentiment classification system.
Methodology	This study utilizes a methodology that includes generating synthetic data to augment the original datasets, designing the input representation, detecting aspect categories, performing a multi-label sentiment classification, and representing sentiment in a four-class sentiment classification.
Contribution	This study provides an investigation of the four-class sentiment classification (positive, negative, neutral, and conflict) and proposes a BERT-based method to identify aspects with conflict sentiment in ABSA. Moreover, it also evaluates Llama 3 for generating synthetic data to address the issues related to data scarcity and imbalanced datasets in the research on four-class sentiment classification in ABSA. The validation of the proposed model on the SemEval-2014 restaurant domain dataset shows an improvement in conflict sentiment accuracy compared to baselines.
Findings	The investigation of the four-class sentiment classification task in ABSA demonstrates that identifying conflict sentiment is challenging for several reasons. Among them are (1) the lack of a public dataset for this research; (2) the small amount of data with conflict labels in the available dataset resulting in an imbalanced dataset; (3) conflict sentiment is a complex sentiment containing both positive and negative sentiments; and (4) conflict sentiments are usually expressed in long and complicated sentences and involve implicit aspects. Our solution to these challenges involved generating synthetic data using Llama 3 and designing a BERT-based model on multi-label aspects for identifying aspect with conflict sentiment. The experimental results demonstrate that our proposed method outperforms previous methods in identifying the fourth sentiment in four-class sentiment classification, i.e., aspects with conflict sentiment.
Recommendations for Practitioners	Most existing ABSA models with four-class sentiment classification are conducted for product reviews (mostly in the restaurant domain) and in high-resource languages (mainly in English). Therefore, users may need to make some adjustments to different domains and languages.
Recommendations for Researchers	Due to the limited availability of datasets for research in aspect-based sentiment analysis with four-class sentiment classification, it is important to urgently develop extra supporting datasets.
Impact on Society	Aspect-based sentiment analysis (ABSA), which employs a four-class sentiment classification (positive, negative, neutral, and conflict), provides a comprehensive analysis about the aspects (or target of opinion) and their sentiment. It will help us understand the sentiment analysis problem better. By providing more accurate sentiment through aspect-based sentiment analysis with four-class sentiment classification, this study can better assist individuals, organizations, or

companies in gaining a view or an opinion about any product, service, or candidate in an electoral vote.

Future Research — Future research on aspect-based sentiment analysis could evaluate other open-source large language models (LLMs), such as Gemma, Mixtral, etc., for generating synthetic data and evaluating the model across various domains and languages. Furthermore, future research could also utilize the LLMs to perform ABSA tasks, such as aspect term extraction, aspect category detection, and sentiment polarities, through fine-tuning the LLMs.

Keywords — aspect-based sentiment analysis, four-class sentiment classification, conflict sentiment, pre-trained language models, large language models

INTRODUCTION

People's reviews or opinions about something are important in influencing every decision. We often look at others' opinions and experiences regarding a product, service, or candidate before purchasing a product, choosing a service, or voting in an election. People's opinions not only affect individuals but also influence organizational decisions. Currently, the rapid advancement of social media applications has made a vast number of people's opinions publicly available and easily accessed. This is the strong motivation for research in sentiment analysis.

Sentiment analysis, often referred to as opinion mining, is one of the popular fields of study in natural language processing (NLP). Sentiment analysis is analyzing individual's opinions, sentiments, evaluations, appraisals, attitudes, and emotions regarding objects such as products, services, organizations, individuals, issues, events, topics, etc. (B. Liu, 2012). It encompasses a wide range of applications and is found in most every domain, from business to social.

In contrast to general sentiment analysis (document- and sentence-level), which considers the same sentiment across an entire document or sentence, aspect-level or aspect-based sentiment analysis predicts the sentiment polarity towards certain targets of an opinionated sentence. The basic idea of an aspect-based sentiment analysis (ABSA) system is that an opinion consists of sentiment and a target. ABSA is sentiment analysis at the phrase or word level. It identifies aspects and predicts their associated sentiment polarity. Aspect in ABSA refers to the word or phrase that defines an opinion in a given sentence. For instance, let's take a sentence from the SemEval-2014 datasets (Pontiki et al., 2014) as an example: "Great food but the service is dreadful." This sentence encompasses two distinct aspects, i.e., food and service, each expressing a different sentiment polarity. The sentiment polarity towards the food aspect is positive, while the sentiment polarity about the service aspect is negative.

Numerous methods have been proposed for solving ABSA tasks. The earliest study on ABSA relied on feature engineering, followed by neural network-based and deep learning methods. More recently, studies with language models (LMs) have been successfully employed to solve ABSA tasks. However, most research only focuses on 2-class or 3-class sentiment classification. The 2-class sentiment classification categorizes sentiment as either positive or negative, while the 3-class sentiment classification classifies sentiment into positive, negative, and neutral. Both 2-class and 3-class sentiment classification exclude conflict sentiment in their classification. Ignoring conflict sentiment causes the classification to be less accurate. An aspect has a conflict sentiment if it expresses both positive and negative sentiments at the same time. For example, the review sentence "The atmosphere was nice, but it was a little too dark." expresses conflict sentiment towards the aspect "atmosphere", as the term "nice" conveys a positive sentiment, while the term "little too dark" conveys a negative sentiment. Classifying the aspect "atmosphere" as positive will result in an inaccurate opinion since the aspect "atmosphere" also has a negative opinion. Likewise, classifying the aspect "atmosphere" as negative

will result in an incorrect opinion since it also has a positive opinion. Adding a conflict sentiment class alongside the positive, negative, and neutral sentiment classes, will produce a more precise assessment than a sentiment analysis utilizing a two- or three-class sentiment classification system. It will better help people find the strengths and weaknesses or the advantage and disadvantage of a product, service, or candidate in an electoral vote.

The limited research on aspect-based sentiment analysis, which includes conflict sentiments alongside positive, negative, and neutral sentiments, stems from several factors. The absence of a sufficient training dataset makes it challenging to include conflict sentiment into the sentiment classification task, particularly in ABSA. The number of aspects with conflict sentiment data in the popular dataset that is widely used for ABSA research is very small, which leads to an imbalanced dataset. Moreover, conflict sentiment is a combination of positive and negative sentiment, making its identification more difficult. Furthermore, conflict sentiments are usually expressed in long and complicated sentences and involve implicit aspects that may require attention.

This study investigates the challenge in research of four-class sentiment classification that includes conflict sentiment in aspect-based sentiment analysis and proposes a model to address the issue. We utilize a methodology that includes designing the input representation, detecting aspect categories, performing a multi-label sentiment classification, representing sentiment in a four-class sentiment classification, and generating synthetic data to augment the original datasets. We propose a model that utilizes a pre-trained language representation model (BERT) with a multi-label aspect. BERT (Devlin et al., 2019), which stands for Bidirectional Encoder Representations from Transformers, is a popular language representation model that helps computers understand and work with human language better. Our proposed model focuses on identifying aspects with conflict sentiment in review sentences. It builds upon earlier research on a BERT-based model with dual-sentiment aspects (Nuryani et al., 2022). The method uses BERT for the sentence-pair classification task by generating pseudo-sentences from aspects and pairing them with review sentences for input representations, as inspired by Sun et al. (2019). Then, we fine-tune the pre-trained BERT to perform multi-label classification and translate the result into a four-class sentiment classification.

Moreover, we also utilize Llama 3 to generate synthetic data. Llama 3, which stands for Large Language Model Meta AI version 3, is an open-source large language model (LLM) developed by Meta. Large language models (LLMs) are types of generative artificial intelligence (Generative AI) that can understand, process, and generate human-like language. LLMs focus specifically on tasks involving natural language, like text generation and understanding. We generate synthetic data with Llama 3 to address the problems associated with data scarcity and imbalanced datasets in the research of four-class sentiment classification of ABSA, where conflict sentiment is included. Synthetic data is artificial data that replicates real-world data. In this study, we evaluate the few-shot prompting of Llama 3. The prompt has four components: role-play (Z. Li et al., 2023) and task specifications for setting the right context, generation conditionals for specifying the desired data type, and in-context demonstrations (Long et al., 2024) for providing implicit human guidance.

Specifically, this paper offers the following contributions:

1. We evaluate open source LLMs, Llama 3, to generate synthetic data for supporting research in the four-class sentiment classification of ABSA, which includes conflict sentiment.
2. We propose a BERT-based model with multi-label aspects for identifying aspects with conflict sentiment in ABSA.
3. We verify our model using the restaurant domain of SemEval-2014 datasets, and the experiment results demonstrate an improvement over the baselines.

We organize the structure of the paper as follows. The next section provides the related works on aspect-based sentiment analysis (ABSA), the BERT-based model for ABSA, and LLM-based data augmentation. Next, we describe our proposed method, which involves synthetic data generation and

a BERT-based model with multi-label aspects. The following section presents and discusses the experiment and its results. The final section presents the paper's conclusion, limitations, and the direction for future research.

RELATED WORKS

ASPECT-BASED SENTIMENT ANALYSIS (ABSA)

Aspect-based sentiment analysis (ABSA), also known as feature-based opinion mining, is a phrase- and word-level sentiment classification (B. Liu, 2012). Different from document- and sentence-level sentiment analysis, which identifies overall sentiment polarity on the entire sentence or document, ABSA will predict the sentiment polarity on certain aspects of an entity. An entity can be a specific characteristic or property of a service or product, such as the price, size, ambience, food, etc.

There are four elements in ABSA (Zhang et al., 2022):

1. The aspect term, or the opinion target, is a phrase or word that is specifically mentioned in the review sentence. For example, consider the following review sentence: "This is a nice cozy place with good pizza." The aspect terms for this review are "place" and "pizza."
2. The aspect category specifies a distinct aspect of an entity for each particular domain. In the above example, "ambience" and "food" are the aspect categories for the restaurant domain.
3. The opinion term is a phrase or word used by an opinion holder to express its sentiment towards a certain target. The opinion terms for the above example are "cozy" and "good."
4. The sentiment polarity refers to the orientation of the sentiment toward a specific aspect term or aspect category. In the above example, the sentiment polarity for the aspect term "place" with the aspect category "ambience" is "positive," and for the aspect term "pizza" with the category "food" is also "positive."

The main tasks in ABSA that have been studied extensively by researchers are the aspect extraction and the aspect sentiment classification task (B. Liu, 2012). The aspect extraction task will identify aspects to be evaluated, and the aspect sentiment classification task will decide the opinions or sentiment on different aspects into positive, negative, neutral, or conflict. There are similar names to define ABSA tasks in other references. For example, opinion target extraction (OTE) or aspect term extraction (ATE) refers to the process of identifying aspect terms; aspect category detection (ACD) detects or identifies aspect categories; sentiment polarity (SP) or aspect sentiment classification (ASC) predicts the sentiment on aspects; aspect category sentiment analysis (ACSA) evaluates sentiment on aspect categories; and aspect term sentiment analysis (ATSA) predicts sentiment on aspect terms (Jiang et al., 2019; Pontiki et al., 2014, 2016).

Early works on ABSA tasks are feature engineering-based models. There are methods based on frequent nouns and noun phrases, exploiting opinion and target relations, using supervised learning, and using topic modelling for aspect extraction tasks. For the aspect sentiment classification tasks, supervised learning and lexicon-based approach are employed (B. Liu, 2012). Then, neural networks and deep learning models were introduced to enhance the performance of ABSA tasks. The use of pre-trained word embeddings like Word2Vec (Mikolov et al., 2013) and Glove (Pennington et al., 2014) on a deep learning-based model showed effectiveness in solving ABSA tasks compared to the early feature-based models. Recently, pre-trained language models (PLMs) such as BERT (Devlin et al., 2019), ELMo (Peters et al., 2018), RoBERTa (Y. Liu et al., 2019), BART (Lewis et al., 2020), and T5 (Raffel et al., 2019) have been extensively exploited for further improving the performance on NLP tasks, including ABSA. Furthermore, the emergence of large language models (LLMs) such as ChatGPT (OpenAI, 2023) and Llama (Touvron et al., 2023) have shown the widespread use and deployment on numerous generative NLP applications, making it interesting to investigate how LLMs perform on ABSA.

BERT ON ABSA

BERT (Bidirectional Encoder Representation from Transformers) is one of the most widely utilized pre-trained language models (PLMs). A pre-trained language model (PLM) is an advanced type of language model (LM), trained on vast datasets. A language model (LM) is a probabilistic representation of a natural language that predicts words in a sentence. It uses many statistical and probabilistic techniques to analyze patterns of word occurrence and prediction from unannotated text data (Hoang et al., 2019). We can fine-tune PLMs for specific tasks like question answering, sentiment analysis, text generation, text summarization, etc. BERT (Devlin et al., 2019) is a PLM that generates deep two-way representations of unlabeled text by using both left and right context across all layers. BERT has two steps in its framework: pre-training, where BERT is trained on large unlabeled data, and fine-tuning, where BERT is first initialized with pre-trained parameters and then fine-tuned using labeled data from specific tasks by adding task-specific layers. BERT is a multi-layer bidirectional Transformer encoder that has the same architecture as the original Transformer (Vaswani et al., 2017). Figure 1 is the architecture of the original Transformer.

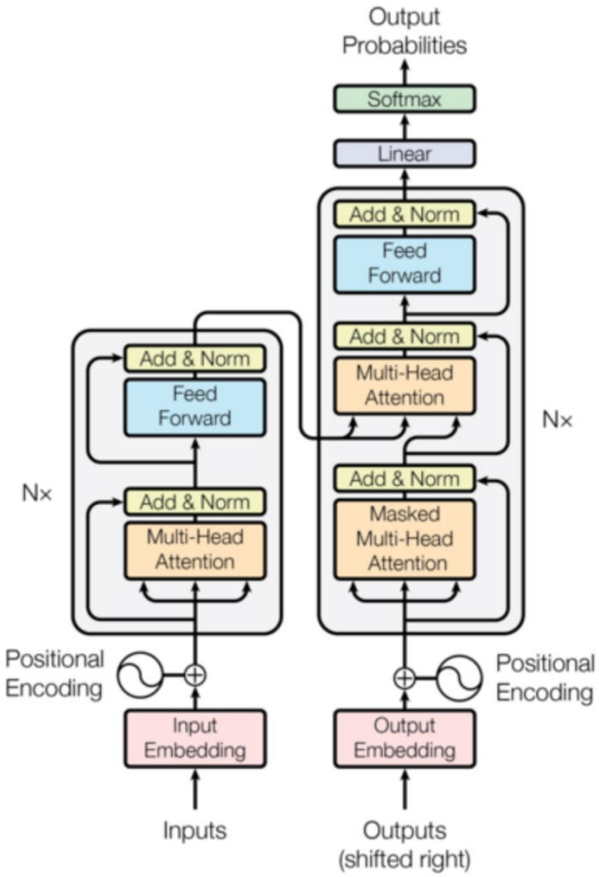

Figure 1. The architecture of Transformer (Vaswani et al., 2017)

The use of BERT has shown a significant improvement on a wide range of NLP tasks, including ABSA. Combining BERT and capsule network (CapsNet-BERT) to do aspect sentiment polarity was investigated by Jiang et al. (2019). To obtain deep representations with pre-trained BERT, the CapsNet-BERT takes aspect (category or term) and review sentences as input and then combines them into capsule layers to do ACSA and ATSA. Hoang et al. (2019) used BERT and fine-tuned it to a

sentence pair classification model to solve out-of-domain ABSA for aspect classification. They proposed a model to predict the aspect related or unrelated to a text, a model to find a relation between aspect and a text to predict the sentiment, and a unified model to classify both aspect and sentiment.

Directly fine-tuning BERT on the end task, with limited tuning data, presents challenges related to domain and task awareness. Therefore, Xu et al. (2019) proposed the BERT Post-Training (BERT-PT) technique as a solution to this issue and evaluated it on ATE and ATSA tasks. Similar work, Rietzler et al. (2020) proposed BERT-ADA, which involves combining a pretrained BERT model fine-tuning on domain-specific data with different end-task training, and then evaluating it on in-domain and cross-domain data.

In review sentences, the words representing aspects and sentiments have positional dependence and are typically located close to each other. It motivated Marcacini and Silva (2021) to propose ABSA using BERT with Disentangled Attention (ABSA-DeBERTa) to predict ATSA. A review sentence may contain multiple aspect categories and express different sentiments toward those aspects. Directly using the given aspect category to identify corresponding sentiment words may result in a mismatching between the sentiment words and the aspect categories. It occurs when an unrelated sentiment word is semantically relevant to the given aspect category. So, to avoid this issue, Y. Li et al. (2020) proposed a Sentence Constituent-Aware Network (SCAN), a graph attention-based model, and combined it with BERT to solve the ACSA.

Sun et al. (2019) proposed a method that utilize BERT to solve ACD and ACSA tasks. This method generates auxiliary sentences from the aspect categories and then convert them into a sentence-pair classification task, similar to those in question answering (QA) and natural language interface (NLI). This method got improvement from auxiliary sentence generation that expanded the dataset for training and its effective transformation of ABSA problems into sentence pair classification tasks. J. Liu et al. (2021) investigate ABSA generation method by casting ACD and ACSA tasks into language generation tasks. They transform ACD and ACSA into sequence-to-sequence (seq2seq) tasks, in which the encoder processes the input sentences, and the decoder produces natural language sentences.

Simmering and Huoviala (2023) proposed a study on evaluating large language models (LLMs) to perform ABSA tasks. They investigated the performance of GPT-4 and GPT-3.5 in zero-shot, few-shot, and fine-tuned settings for solving ATE and ATSA tasks. Their fine-tuned GPT-3.5 on SemEval-2014 Task 4 achieves a state-of-the-art result, but with more model parameters and increased inference cost. Wang et al. (2024) also propose few-shot LLM prompting for ATE and ATSA tasks using the in-context learning (ICL) paradigm. Zhang et al. (2024) investigated LLMs on various sentiment analysis tasks and discovered that while LLMs perform well in general or conventional sentiment analysis, they struggle in more complex sentiment analysis tasks like ABSA and MAST (multifaceted analysis of subjective texts).

Many languages model-based approaches have been proposed to tackle ABSA tasks. Most of them are limited to 2- or 3-class classification and do not incorporate conflict sentiments in their classification tasks (Jiang et al., 2019; Y. Li et al., 2020; J. Liu et al., 2021; Marcacini & Silva, 2021; Rietzler et al., 2020; Simmering & Huoviala, 2023; Wang et al., 2024; Xu et al., 2019; Zhang et al., 2024). One study that carried out 4-class classification for ABSA was by Sun et al. (2019). They generate auxiliary sentences from aspects, pair them with review sentences and then transform them into a BERT-based model for sentence pair classification tasks. Tan et al. (2019) examined 4-class ABSA approach, focusing on recognizing conflict sentiment in ABSA. They proposed the Dual Attention-based GRU (D-AT-GRU), a multi-label classification model with a dual attention mechanism. This model takes review sentences and aspects as inputs, transforms them into word and aspect embeddings, then extracts aspect-related text features using RNN and the dual attention mechanism, performs a 2-label classification, and lastly transforms them into a 4-class classification.

P. Li et al. (2022) proposed Dual Multi-head Attention Edge Convolutional (D-MA-EGCN), a graph neural network model designed to improve the D-AT-GRU (Tan et al., 2019). To encode the sentences, they combine BERT with the edge graph convolutional neural network (EGCN) to extract the relationship between the aspect word and the sentiment word. Sun et al. (2019) reported a significant improvement in the BERT-auxiliary sentences model's ability to classify four-class sentiment on ABSA. Tan et al., (2019) also reported a way to make complicated conflict sentiments easier to understand. This inspired Nuryani et al. (2022) to utilize BERT to study how to identify conflict sentiments in ABSA. They proposed a BERT-based method with dual-sentiment aspects. To do this, they assigned each aspect a value for both positive and negative sentiment, paired them with review sentences to represent input in a sentence pair classification task, performed a multi-label classification task, and sum up the results into a four-class sentiment classification.

LLMs-Based Data Augmentation

As previously mentioned, most proposed methods on ABSA exclude conflict sentiment for some reason. One of the reasons is the insufficiency and imbalance of datasets due to the limited number of aspects with conflict sentiment in the datasets. Since data annotation often requires significant time and cost, strategies such as data augmentation and synthetic data generation can provide viable solutions to address this challenge. Data augmentation is a technique employed in machine learning and deep learning to create new data from existing data. Synthetic data generation serves as a supplementary technique to data augmentation. It artificially generates data that mimics the real data. Data augmentation and synthetic data generation are especially useful when datasets for training are small or imbalanced.

Many studies have proposed methodologies for augmenting existing annotated data or generating synthetic data, either with or without a few annotated data samples, with the aim of expanding the training dataset. Simple text data augmentation techniques include replacing words with WordNet synonyms or neighbors in the embedding space, a combination of replacing, inserting, and merging with a pre-trained masked language model, combining word insertions, substitutions or deletions, a combination of substituting, deleting, inserting, and swapping adjacent characters, the backtranslation method, back transcription approach, etc. (Morris et al., 2020).

Recent advancements in large language models (LLMs) have sparked various studies on generating synthetic data and augmenting text data using LLMs to support NLP research (Dai et al., 2023; Z. Li et al., 2023; R. Liu et al., 2024; Long et al., 2024; Samuel et al., 2024; Ubani et al., 2023). Dai et al. (2023) propose a text data augmentation method utilizing ChatGPT, referred to as AugGPT. AugGPT uses few-shot prompting to rephrase each sentence in the training dataset into several similar sentences. The AugGPT shows superior performance on text classification tasks. Ubani et al. (2023) proposed the ZeroShotDataAug, a zero-shot prompting ChatGPT for data augmentation, as another ChatGPT-based data augmentation method. The zero-shot setting can be a promising strategy if any labeled training data are unavailable.

Samuel et al. (2024) evaluate the generating of synthetic data using LLMs for specific NLP tasks. They employ a few-shot setting for GPT-4 to augment existing QA system datasets for replacing human annotators. Z. Li et al. (2023) evaluate synthetic data generation from zero-shot and few-shot LLMs (GPT-3.5-Turbo) settings on various text classification tasks. They found that synthetic data is more effective for tasks with low subjectivity, like AG's news topic classification, relation classification, IMDB reviews, and SMS spam datasets, than for tasks with high subjectivity.

PROPOSED METHOD

This section describes our proposed method in detail. We divide the proposed method into two main stages: the synthetic data generation and the BERT-based with multi-label aspects.

SYNTHETIC DATA GENERATION

Our goal in generating synthetic data is to address issues related to dataset limitations for research in the four-class sentiment classification of ABSA, which include positive, negative, neutral, and conflict. One of the well-known public datasets for research in this field is the SemEval-2014 restaurant domain. The limited number of aspects exhibiting conflict sentiments leads to an imbalance in these datasets. Therefore, we generate sentences containing aspects with conflict sentiment to improve the balance level of the datasets. We utilize Llama-3, an open-source LLM developed by Meta.

In generating synthetic data, we ensure that the synthetic data produced meets the main requirement for high-quality data, i.e., diversity and faithfulness (Long et al., 2024). We divide the process of synthetic data generation into two steps: data generation and data curation. For the first step, the design of data generation ensures the diversity of the generated synthetic data. We achieve this by carefully designing the prompt. In LLMs, a prompt functions as an instruction or a question, guiding the model toward a specific response. We type the input text to tell the model what we need. Based on the availability of real data samples, we can design prompts using two techniques: zero-shot and few-shot prompting. Zero-shot prompting is when we interact with LLMs without providing any examples, and few-shot prompting is if we give a few examples in the prompt. The second step is data curation. We design data curation to guarantee the use of only correct or valid data in the next stage of the research. Figure 2 provides the illustration of the synthetic data generation process in our work.

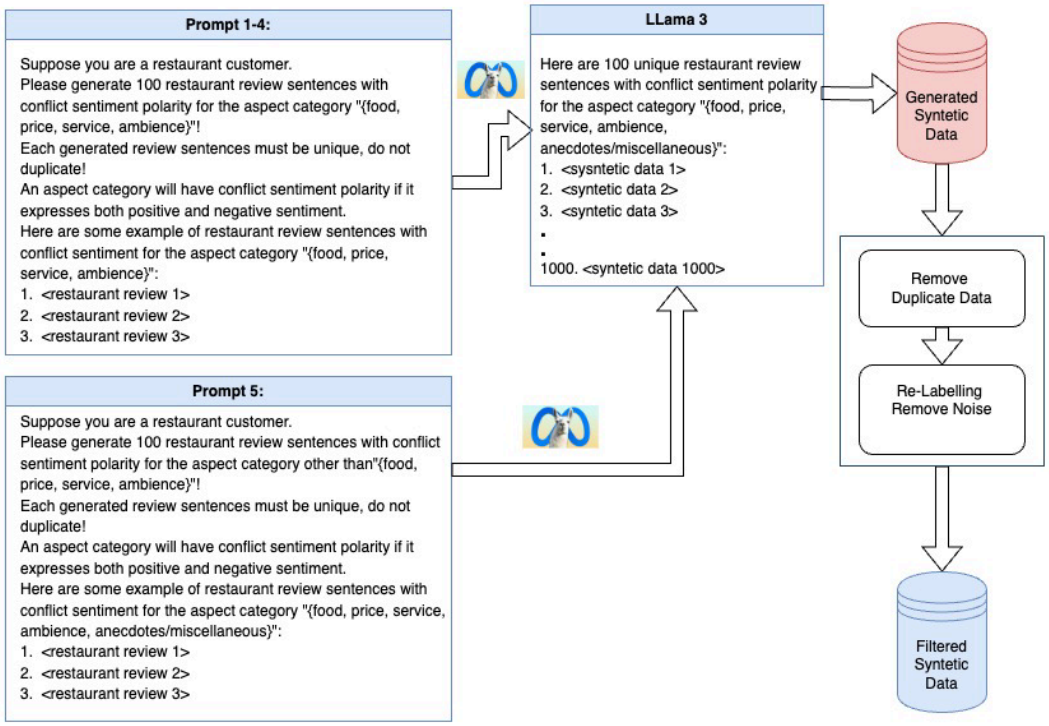

Figure 2. The illustration for synthetic data generation process

Data generation

We use a few-shot prompting technique for data generation, which consists of three elements: task specification, generation conditions, and in-context demonstrations. We intend the task specification to provide the appropriate context for LLMs in generating data. In this task specification, we use role-playing to establish the proper scenario (Z. Li et al., 2023). For the generation conditions, we explicitly describe the desired data type and directly communicate with the LLMs to generate diverse data. For the in-context demonstrations, we provide three instances of real-world data (taken from the original training dataset) to guide LLMs in generating faithful data. Figure 3 illustrates the detailed explanation for the synthetic data generation prompt.

To generate more diverse data, we run the generation process in two rounds for each aspect category. In our work, we use restaurant domains with five aspect categories: <food>, <price>, <service>, <ambience>, and <anecdotes/miscellaneous> (as in SemEval 2014 datasets). We apply the same prompt for the aspect categories: <food>, <price>, <service>, and <ambience>. For the aspect category <anecdotes/miscellaneous>, we specifically use "*other than food, price, service, and ambience*" instead of directly using the term "*anecdotes/miscellaneous*" in the generation conditions to provide the Llama with clearer instructions. This prompt effectively generates a wider range of synthetic data for the aspect category <anecdotes/miscellaneous>.

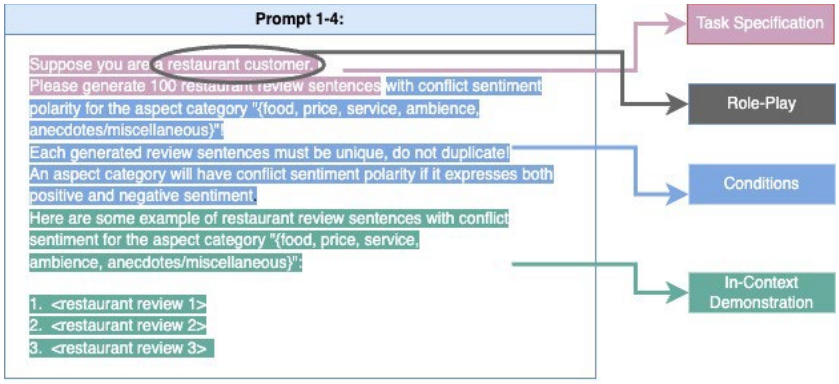

Figure 3. The explanation for synthetic data generation prompt

Data curation

The LLM-generated synthetic datasets often comprise noisy, worthless, or even toxic samples. Therefore, directly employing this dataset may have a negative impact on the model (Long et al., 2024). In our work, the generated dataset from the previous step contains duplicate data, as well as some mislabeled and irrelevant data (noise). Therefore, in the second step of the data generation process, the data curation, we apply two types of filtering: (1) duplicate data removal and (2) manual re-labeling and noise removal. Figure 4 illustrates the data curation process, while the implementation of data curation in our work is detailed in the result and evaluation section (Figure 7).

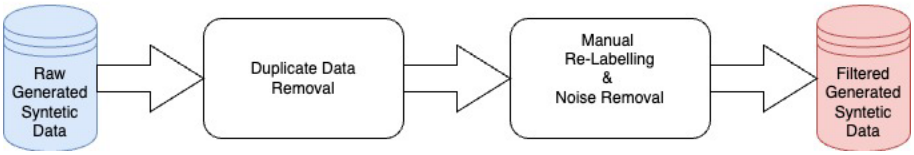

Figure 4. The data curation process

BERT-Based Model with Multi-Label Aspect

We proposed a BERT-based model with multi-label aspects for a sentence-pair classification task. This approach drew inspiration from previous work on BERT-auxiliary sentences on dual-sentiment aspects (Nuryani et al., 2022). We design the model to perform 4-class sentiment classification in ABSA, with a focus on better identification of conflict sentiment.

A multi-label aspect is defined as an aspect with multiple labels. This proposed method assigns a value to each aspect for the labels positive, negative, and none. Aspects with positive sentiment will have value <yes/1> for <aspect - positive>, while <aspect - negative> and <aspect - none> will have value <no/0>. If the aspect has a negative sentiment, then <aspect - negative> is <yes/1>, while <aspect - positive> and <aspect - none> are <no/0>. For the neutral aspect, both <aspect - positive> and <aspect - negative> are <no/0>, while for the conflict aspect, both are <yes/1> with <aspect - none> being <no/0>. The <aspect - none> will be <yes/1> if the aspect is unrelated to the given review sentence. Table 1 shows the conversion of sentiment to a multi-label aspect.

In the proposed method, we jointly evaluate the model for two ABSA subtasks: the aspect category detection (ACD) and the aspect category sentiment analysis (ACSA). Figure 5 depicts the architecture of the model.

Table 1. The conversion of sentiment on multi-label aspect

ORIGINAL LABEL /SENTIMENT	ASPECT_ POSITIVE	ASPECT_ NEGATIVE	ASPECT_ NONE
Unrelated Aspect	0	0	1
Neutral	0	0	0
Negative	0	1	0
Positive	1	0	0
Conflict	1	1	0

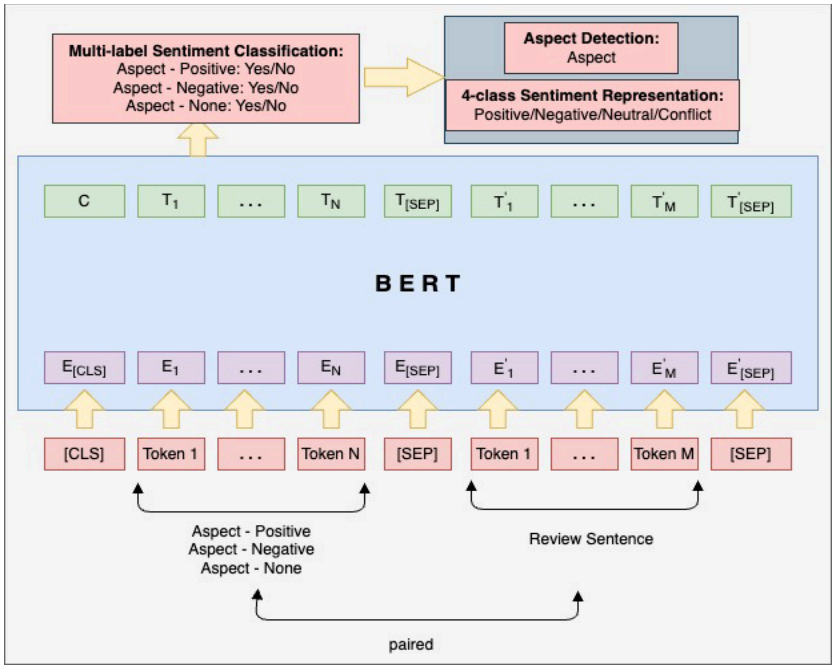

Figure 5. The architecture for BERT-based model with multi-label aspects

As illustrated in Figure 5, the model takes pseudo-sentences for multi-label aspects (aspect – positive, aspect – negative, and aspect – none) paired with review sentences as an input, which are then represented in a single token sequence. In this model, we employ BERT-NLI-B (Sun et al., 2019) to generate pseudo-sentences from aspects. We then pair these pseudo-sentences with review sentences as input and transform them into a sentence-pair classification task, similar to natural language inference (NLI). The formula for pseudo-sentence generation is "{aspect_category} - {positive, negative, none}" with label {Yes/1} if the condition is matched or {No/0} otherwise.

For a detailed explanation, consider the following review sentence: "The food was a little burnt but still delicious." This sentence has the label "conflict" for the aspect category "food." The inputs for the model are <food – positive> and <food – negative> both of which have labels <Yes/1> and <No/0> for <service – none>. Each paired with the review sentence "The food was a little burnt but still delicious." The labels <Yes/1> on <food – positive> and <food – negative> indicate that the aspect category "food" has both "positive" and "negative" sentiment, resulting in a conflict sentiment. Other pairs for other aspect categories (the restaurant domain of SemEval datasets have five categories: food, service, price, ambience, and anecdotes), the label for both <aspect_category – positive> and <aspect_category – negative> will be <No/0> while <aspect_category – none> will be <Yes/1>. The detailed explanation is given in Table 2.

Table 2. Example of sentence pair generation for model's input

PSEUDO-SENTENCES	REVIEW SENTENCES	LABEL
food-positive	The food was a little burnt but still delicious.	0
food-negative	The food was a little burnt but still delicious.	0
food-none	The food was a little burnt but still delicious.	1
service-positive	The food was a little burnt but still delicious.	1
service-negative	The food was a little burnt but still delicious.	1
service-none	The food was a little burnt but still delicious.	0
price-positive	The food was a little burnt but still delicious.	0
price-negative	The food was a little burnt but still delicious.	0
price-none	The food was a little burnt but still delicious.	1
ambience-positive	The food was a little burnt but still delicious.	0
ambience-negative	The food was a little burnt but still delicious.	0
ambience-none	The food was a little burnt but still delicious.	1
anecdotes/miscellaneous-positive	The food was a little burnt but still delicious.	0
anecdotes/miscellaneous-negative	The food was a little burnt but still delicious.	0
anecdotes/miscellaneous-none	The food was a little burnt but still delicious.	1

The input representations for the BERT-based model are formulated as "[CLS] pseudo_sentence [SEP] review_sentence [SEP]". [CLS] is a special classification token for the first token of every sequence and [SEP] is a special token to differentiate the pseudo-sentence or review sentence. The representation of input for a given token is calculated by summing the associated token, segment, and position embeddings.

In fine-tuning procedures, the final hidden state (i.e., output of the transformer) of the first token [i.e., CLS] is fed into the classification layer to perform the multi-label sentiment classification. The initial output from the classification layer is multi-label sentiments for multi-label aspects, which is then translated into detected aspect and four-class sentiment classification (positive, negative, neutral, conflict) through aspect category detection and four-class sentiment representation tasks.

EXPERIMENT AND RESULTS

DATASETS

We evaluate the proposed model, a BERT-based model with multi-label aspects, on the restaurant domain of the SemEval-2014 datasets. The dataset has 3044 sentences for training data and 800 sentences for testing data. The datasets have four sentiment labels (positive, negative, neutral, and conflict) and five aspect categories (food, service, price, ambience, and anecdotes/miscellaneous). The original datasets also include aspect terms and aspect term sentiment polarities. Since we evaluate our proposed model for the aspect category detection (ACD) and the aspect category sentiment classification (ACSA) task of ABSA, we do not use aspect terms and aspect term sentiment polarities. We then modified these datasets by adding synthetic data generated by Llama. The outcome is a dataset that includes 3370 sentences and 4038 aspects. Table 3 displays the detailed statistics for the datasets.

Table 3. The statistics of SemEval-2014 restaurant domain

DATASET	POSITIVE	NEGATIVE	NEUTRAL	CONFLICT
Training data (original) 3044 sentences 3712 aspects	2176 (58.7%)	839 (22.6%)	501 (13.5%)	196 (5.3%)
Training data (after modification) 3370 sentences 4038 aspects	2176 (53.9%)	839 (20.8%)	501 (12.4%)	522 (12.9%)
Testing data 800 sentences 1025 aspects	657 (64.1%)	222 (21.7%)	94 (9.2%)	52 (5.1%)

IMPLEMENTATION DETAILS

For synthetic data generation, we use Llama3 with 70 B parameters size, known as llama3_70b, and run the model in Groq AI Inference. We set the temperature to 0.7 to control the randomness of the model's output. We generate 100 sentences for each aspect category and run the generation process in two rounds, so we have 1000 sentences in the first-step generation. For data curation, we apply duplicate data removal and get 542 sentences, or 54.2% unique sentences. After applying manual relabeling and noise removal, we obtain 326 sentences, or 32.6%, in the final step of synthetic data generation. We then add the synthetic data generated by Llama to the original data from SemEval-2014 for fine-tuning purposes.

For BERT fine-tuning, we follow the hyperparameter settings in BERT-auxiliary sentences (Sun et al., 2019) with some modifications. We utilized $BERT_{BASE}$ with 12 layers of transformer blocks (L), a hidden layer size (H) of 768, 12 self-attention heads (A), and 110 M parameters. We defined the learning rate to 2e-5 and the dropout probability to 0.1. To accommodate the available infrastructure, we defined the batch size to 12, the seed to 21, and the number of epochs to 4. We evaluate the

model using Macro-F1 for the ACD task and accuracy score for the ACSA task. We summarize the implementation details of our work in Table 4.

Table 4. The implementation details summary

STEP	HYPERPARAMETERS	VALUE
Synthetic Data Generation	LLM model	llama3-70b
	Temperature	0.7
	Prompt	Few-shot, 3 examples
	Inference	Groq AI Inference
BERT fine-tuning	BERT model	BERTBASE: 12 layers of transformer blocks (L), 768 of hidden layer size (H) of, 12 self-attention heads (A), 110 M parameters
	Learning rate	2e-5
	Dropout probability	0.1
	Batch size	12
	Seed	21
	Number of epochs	4

RESULTS AND EVALUATION

Table 5 summarizes experimental results of our proposed method. We compare our proposed method with BERT-NLI-B (Sun et al., 2019) as a baseline to show the effectiveness of the designed multi-label aspects. To demonstrate the significance of Llama-generated synthetic data, we also compare it to other augmentation methods, such as embedding-based and pre-trained language models (PLMs)-based text data augmentation.

Table 5. The experiment results on BERT-based model

MODEL	EVALUATION METRICS				
	P	R	MACRO-F1	ACCURACY (%)	
				CONFLICT	OVERALL
BERT-NLI-B	0.929	0.893	0.911	34.62	83.51
BERT-MLA*	0.935	0.897	0.915	44.23	79.90
BERT-MLA-Emb-Aug**	0.932	0.899	0.915	32.69	80.20
BERT-MLA-PLM-Aug***	0.936	0.906	0.921	38.46	80.00
BERT-MLA-Llama-Aug****	0.927	0.899	0.913	**51.92**	82.24

* MLA: multi-label aspects **Emb-Aug: embedding-based augmentation
***PLM-Aug: PLM-based augmentation ****Llama-Aug: Llama-based augmentation

Table 5 shows that our proposed method, designing multi-label aspects for BERT (BERT-MLA), outperforms the baseline in the accuracy of label conflict. The proposed method has improved the

accuracy of conflict labels by 9.61% compared to the baseline method. It demonstrates that designing multi-label aspects and generating pseudo-sentences from the aspects effectively identifies the fourth sentiment on the four-class sentiment classification in ABSA: aspects with conflict sentiments. Assigning each aspect with multi-labels (positive and negative) at the same time will provide the model with clues in simplifying the complexity of conflict sentiment. In our proposed method, we also include the label "none" in the multi-label aspect as an additional label to identify the aspect category.

Moreover, the proposed method with Llama-generated synthetic data achieves the highest accuracy for conflict class. By incorporating only 10.71% of synthetic data for the conflict class, the model improves its accuracy for conflict labels by 17.3% over the baseline and 7.69% over the BERT with multi-label aspects without added synthetic data for data training. The experiment result shows that synthetic data generated by Llama 3 has a significant potential for providing low-resource data to support the model's training.

Llama-based data generation achieves the highest performance compared to other text data augmentation techniques, such as word embedding and PLM-based augmentation. Word embedding and PLM-based data augmentation have been negatively associated with classification performance. Using word embedding and PLM-based augmentation can generate instances with shifted feature spaces, resulting in lower performance on classification tasks (Yang & Li, 2024). The word embedding and PLM-based augmentation can generate out-of-vocabulary words for word substitution and insertion. The augmentation process can alter the meaning of the augmented sentences, leading to adverse interpretations in sentiment analysis tasks. In our experiment, embedding-based augmentation loses 11.54%, while PLM-based augmentation loses 5.77% in the performance on conflict label classification.

To generate more diverse data, we use a single prompt for each aspect category during the synthetic data generation. Specifically for the aspect category "anecdotes/miscellaneous," we use the phrase "other than food, price, service, and ambience" instead of the term "anecdotes/miscellaneous" itself. This scenario generates more diverse data compared to generated data from other aspect categories. We obtained 12.7% high-quality data for the aspect category "anecdotes/miscellaneous" from 1000 total data generated by Llama 3. The aspect category "price" generated the least amount of data, with 1.1%, followed by "food" with 5.1%, "service" with 6.2%, and "ambience" with 7.5%. Figure 6 presents a comparison of the high-quality synthetic data generated by Llama 3 for each aspect category.

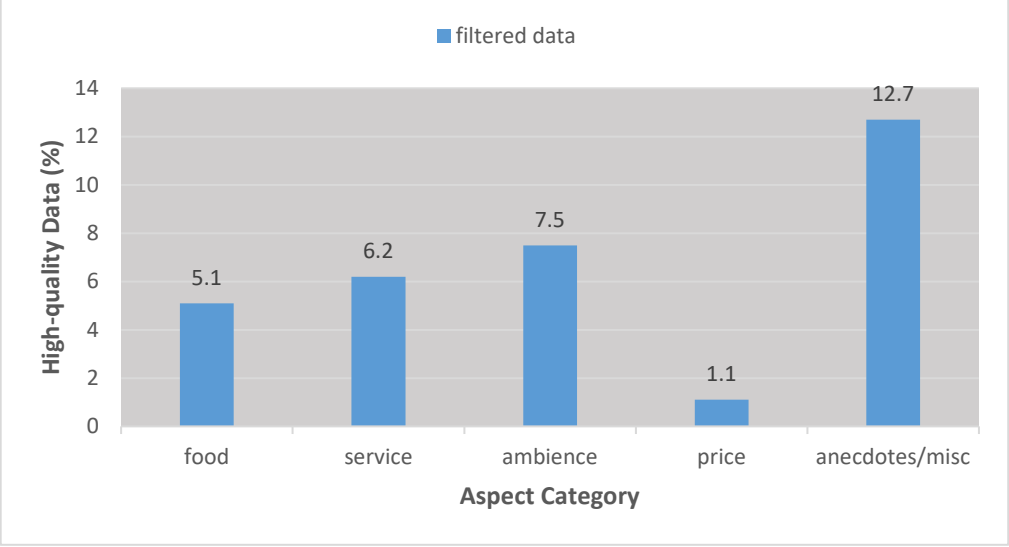

Figure 6. The comparison of high-quality synthetic data for each aspect category generated by Llama 3

The low amount of generated data for the aspect category price might be related to the limited number of components that define its aspects and sentiments. The aspect category price can only be defined by cost, money, payment, or monetary denominations such as dollars, euros, etc. And the sentiment for price is limited to the terminology that refers to cheap or expensive. This is different from other aspect categories that have many components, such as the food, which can be defined through various types of food, menus, portions, etc., and food sentiment can be defined to express taste, portions, menu, food presentation, etc.

The experiment results demonstrate that Llama 3 has a high potential for generating high-quality synthetic data. However, it also generates a significant amount of duplicate data, even at relatively high temperature settings, and directly communicates with the Llama to generate unique data through prompts. We employ two types of data filtering in the data curation process, as described in the proposed method section (Figure 4). In our study, we set the temperature to 0.7 and instructed Llama 3 to generate 2x100 sentences for each aspect category, resulting in 1000 synthetic data. These datasets contain 558 sentences, or 55.8% duplicate data; hence, after filtering to eliminate duplicate data, we retained only 542 sentences, or 54.2% of the unique data. The second filtering involves manual relabeling and noise removal, resulting in just 326 sentences, or 32.6% of high-quality data. Figure 7 illustrates the process of obtaining filtered data from raw datasets containing 1000 sentences generated by Llama 3.

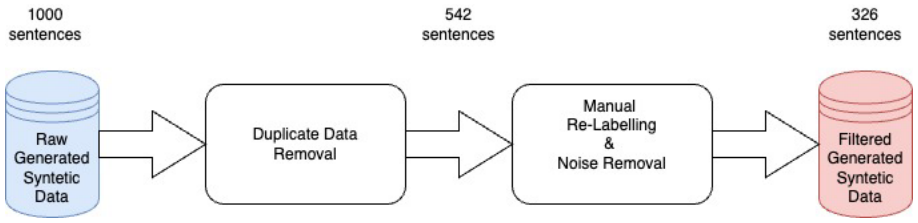

Figure 7. The process of obtaining filtered data

The proposed method improves performance in conflict sentiment classification by 17.3%, but overall performance drops by 2.3%. There were 15 instances with positive sentiment labels, but the model predicted them as conflict. And out of the 15 sentences, 8 can indeed be classified as conflict. For example, the sentence in the testing data "Although small, it has beautiful ambience, excellent food (the catfish is delicious - if ya don't mind it a lil salty) and attentive service" has positive labels for the aspect categories "ambience," "food," and "service." The aspect category "ambience" can be classified as conflict based on the sentence clause "Although small, it has beautiful ambience." Similarly, the sentences "The food is all-around good, with the rolls usually excellent and the sushi/sashimi not quite on the same level," "The ambiance was fine, a little loud but still nice and romantic," and "Cool atmosphere but such a let down" all have positive labels for the aspect category of food and ambience, respectively. Given these facts, additional analysis of the dataset annotations may be required, and we believe that the performance of our proposed model can still be improved.

Furthermore, like the training data, the number of conflict labels in the testing data on the SemEval-2014 restaurant domain dataset is also very small, at only 5.1%. This confuses model evaluation because maximizing the accuracy of the conflict label does not maximize the overall performance of the model. Therefore, the development of datasets to support this research, four-class sentiment classification in aspect-based sentiment analysis, is urgently needed.

CONCLUSIONS, LIMITATIONS, AND FUTURE RESEARCH

In this paper, we proposed a BERT-based model on multi-label aspects to conduct four-class classification in aspect-based sentiment analysis, focusing on identification of conflict sentiment. We improved the previous work by designing multi-label aspects, with each aspect assigned a value for positive and negative sentiments. Then, we generate pseudo-sentences from those aspects that are then paired with review sentences for input and transform them to a sentence-pair classification task. Furthermore, we also generate synthetic data with Llama 3, an open-source large language model (LLM) developed by Meta, to address the issues related to the data scarcity and imbalanced datasets in the research of four-class sentiment classification, aspect-based sentiment analysis. The experimental results on the SemEval-2014 datasets with the restaurant domain show that our proposed method outperforms earlier methods in identifying the fourth sentiment in four-class sentiment classification, i.e., aspects with conflict sentiment.

The proposed model features two significant enhancements: the implementation of multi-label aspects for BERT and the synthetic data generation using Llama 3. The design of multi-label aspects enables the model to recognize intricate manifestations of conflict sentiments. This shows that designing multi-label aspects for a BERT-based model is effectively identifying the conflict sentiment in ABSA. Llama-generated synthetic data provides a solution to the limitations and imbalances of the datasets. This indicates that synthetic data generation with Llama3 has significant potential to support research with low-resource datasets.

Although our proposed model has shown promising results in identifying conflict sentiment in ABSA, it also has certain limitations. For example, we solely use Llama 3 with 70 B parameters as the LLM to generate synthetic data. We also evaluate the model on specific datasets, i.e., datasets with restaurant domain and English language. Future research should take these into consideration and address the related issues. The promising direction for future research involves evaluating other open-source LLMs, such as Gemma and Mixtral, for generating synthetic data, and evaluating the model across various domains and languages. In addition, we could utilize the LLMs to perform ABSA tasks, such as aspect term extraction, aspect category detection, and sentiment polarities, through fine-tuning the LLMs.

REFERENCES

Dai, H., Liu, Z., Liao, W., Huang, X., Cao, Y., Wu, Z., Zhao, L., Xu, S., Liu, W., Liu, N., Li, S., Zhu, D., Cai, H., Sun, L., Li, Q., Shen, D., Liu, T., & Li, X. (2023). AugGPT: Leveraging ChatGPT for text data augmentation. *arXiv preprint*. http://arxiv.org/abs/2302.13007

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint*. http://arxiv.org/abs/1810.04805

Hoang, M., Bihorac, O. A., & Rouces, J. (2019). Aspect-based sentiment analysis using BERT. *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, 187–196. https://aclanthology.org/W19-6120/

Jiang, Q., Chen, L., Xu, R., Ao, X., & Yang, M. (2019). A challenge dataset and effective models for aspect-based sentiment analysis. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 6279–6284. https://doi.org/10.18653/v1/D19-1654

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., & Zettlemoyer, L. (2020). BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7871–7880. https://doi.org/10.18653/v1/2020.acl-main.703

Li, P., Chang, W., Zhou, S., Xiao, Y., Wei, C., & Zhao, R. (2022). A conflict opinion recognition method based on graph neural network in aspect-based sentiment analysis. *2022 5th International Conference on Data Science and Information Technology (DSIT)*, 1–6. https://doi.org/10.1109/DSIT55514.2022.9943870

Li, Y., Yin, C., & Zhong, S. (2020). Sentence constituent-aware aspect-category sentiment analysis with graph attention networks. *arXiv preprint*. http://arxiv.org/abs/2010.01461

Li, Z., Zhu, H., Lu, Z., & Yin, M. (2023). Synthetic data generation with large language models for text classification: potential and limitations. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 10443–10461. https://doi.org/10.18653/v1/2023.emnlp-main.647

Liu, B. (2012). *Sentiment analysis and opinion mining*. Morgan & Claypool. https://doi.org/10.1007/978-3-031-02145-9

Liu, J., Teng, Z., Cui, L., Liu, H., & Zhang, Y. (2021). Solving aspect category sentiment analysis as a text generation task. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 4406–4416. https://doi.org/10.18653/v1/2021.emnlp-main.361

Liu, R., Wei, J., Liu, F., Si, C., Zhang, Y., Rao, J., Zheng, S., Peng, D., Yang, D., Zhou, D., & Dai, A. M. (2024). Best practices and lessons learned on synthetic data. *arXiv preprint*. http://arxiv.org/abs/2404.07503

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint*. https://doi.org/10.48550/ARXIV.1907.11692

Long, L., Wang, R., Xiao, R., Zhao, J., Ding, X., Chen, G., & Wang, H. (2024). On LLMs-driven synthetic data generation, curation, and evaluation: A survey. *arXiv preprint*. http://arxiv.org/abs/2406.15126

Marcacini, R., & Silva, E. (2021, July 24). Aspect-based sentiment analysis using BERT with disentangled attention. *LatinX in AI at International Conference on Machine Learning 2021*. https://doi.org/10.52591/lxai2021072416

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint*. https://doi.org/10.48550/ARXIV.1301.3781

Morris, J., Lifland, E., Yoo, J. Y., Grigsby, J., Jin, D., & Qi, Y. (2020). TextAttack: A framework for adversarial attacks, data augmentation, and adversarial training in NLP. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 119–126. https://doi.org/10.18653/v1/2020.emnlp-demos.16

Nuryani, N., Purwarianti, A., & Widyantoro, D. H. (2022). Identification of conflict opinion in aspect-based sentiment analysis using BERT-based method. *Proceedings of the 2022 International Conference on Computer, Control, Informatics and Its Applications*, 276–280. https://doi.org/10.1145/3575882.3575935

OpenAI. (2023). GPT-4 technical report. *arXiv preprint*. https://doi.org/10.48550/ARXIV.2303.08774

Pennington, J., Socher, R., & Manning, C. (2014). GloVe: Global vectors for word representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543. https://doi.org/10.3115/v1/D14-1162

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 2227–2237. https://doi.org/10.18653/v1/N18-1202

Pontiki, M., Galanis, D., Papageorgiou, H., Androutsopoulos, I., Manandhar, S., AL-Smadi, M., Al-Ayyoub, M., Zhao, Y., Qin, B., De Clercq, O., Hoste, V., Apidianaki, M., Tannier, X., Loukachevitch, N., Kotelnikov, E., Bel, N., Jiménez-Zafra, S. M., & Eryiğit, G. (2016). SemEval-2016 Task 5: Aspect based sentiment analysis. *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, 19–30. https://doi.org/10.18653/v1/S16-1002

Pontiki, M., Galanis, D., Pavlopoulos, J., Papageorgiou, H., Androutsopoulos, I., & Manandhar, S. (2014). SemEval-2014 Task 4: Aspect based sentiment analysis. *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, 27–35. https://doi.org/10.3115/v1/S14-2004

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2019). Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint*. https://doi.org/10.48550/ARXIV.1910.10683

Rietzler, A., Stabinger, S., Opitz, P., & Engl, S. (2020). Adapt or get left behind: domain adaptation through BERT language model finetuning for aspect-target sentiment classification. *Proceedings of the Twelfth Language Resources and Evaluation Conference,* 4933–4941. https://aclanthology.org/2020.lrec-1.607.pdf

Samuel, V., Aynaou, H., Chowdhury, A. G., Ramanan, K. V., & Chadha, A. (2024). Can LLMs augment low-resource reading comprehension datasets? Opportunities and challenges. *arXiv preprint*. http://arxiv.org/abs/2309.12426

Simmering, P. F., & Huoviala, P. (2023). Large language models for aspect-based sentiment analysis. *arXiv preprint*. http://arxiv.org/abs/2310.18025

Sun, C., Huang, L., & Qiu, X. (2019). Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence. *arXiv preprint*. http://arxiv.org/abs/1903.09588

Tan, X., Cai, Y., & Zhu, C. (2019). Recognizing conflict opinions in aspect-level sentiment classification with dual attention 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)*, 3424–3429. https://doi.org/10.18653/v1/D19-1342

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., & Lample, G. (2023). LLaMA: Open and efficient foundation language models. *arXiv preprint*. https://doi.org/10.48550/ARXIV.2302.13971

Ubani, S., Polat, S. O., & Nielsen, R. (2023). ZeroShotDataAug: Generating and augmenting training data with ChatGPT. *arXiv preprint*. http://arxiv.org/abs/2304.14334

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *arXiv preprint*. https://doi.org/10.48550/ARXIV.1706.03762

Wang, Q., Xu, H., Ding, K., Liang, B., & Xu, R. (2024). In-context example retrieval from multi-perspectives for few-shot aspect-based sentiment analysis. *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024),* 8975–8985. https://aclanthology.org/2024.lrec-main.786/

Xu, H., Liu, B., Shu, L., & Yu, P. (2019). BERT post-training for review reading comprehension and aspect-based sentiment analysis. *Proceedings of the 2019 Conference of the North*, 2324–2335. https://doi.org/10.18653/v1/N19-1242

Yang, H., & Li, K. (2024). BootAug: Boosting text augmentation via hybrid instance filtering framework. *arXiv preprint*. http://arxiv.org/abs/2210.02941

Zhang, W., Deng, Y., Liu, B., Pan, S., & Bing, L. (2024). Sentiment analysis in the era of large language models: a reality check. *Findings of the Association for Computational Linguistics: NAACL 2024*, 3881–3906. https://doi.org/10.18653/v1/2024.findings-naacl.246

Zhang, W., Li, X., Deng, Y., Bing, L., & Lam, W. (2022). A survey on aspect-based sentiment analysis: tasks, methods, and challenges. *arXiv preprint*. http://arxiv.org/abs/2203.01054

AUTHORS

Nuryani received a bachelor's degree in computer science from Gadjah Mada University (UGM) and a master's degree in electrical engineering from Bandung Institute of Technology (ITB). Currently, she is pursuing a doctoral degree from the School of Electrical and Informatics Engineering, Bandung Institute of Technology (ITB). She is also a young researcher at the National Research and Innovation Agency (BRIN)—Indonesia. Her research interests include computer networks, deep learning, and natural language processing.

Rinaldi Munir received bachelor's, master's, and doctoral degrees in informatics engineering from Bandung Institute of Technology (ITB). He is an associate professor at the School of Electrical and Informatics Engineering, Bandung Institute of Technology (ITB), Indonesia. His research interest includes cryptography, image processing, watermarking, numerical methods, and algorithms.

Ayu Purwarianti received a bachelor's and master's degree in informatics engineering from Bandung Institute of Technology (ITB). She received a doctoral degree from Toyohashi University of Technology—Japan. She is an associate professor at the School of Electrical and Informatics Engineering, Bandung Institute of Technology (ITB), Indonesia. She is also a co-founder at Prosa AI and a researcher at the Artificial Intelligence Center - ITB. Her research interest includes machine learning and computational linguistics, with a particular focus on Indonesian computational linguistics, including NLP tools, question answering, text categorization, and information extraction.

Dessi Puji Lestari received a bachelor's degree in informatics engineering from Bandung Institute of Technology (ITB). She received her master's and doctoral degrees from the Tokyo Institute of Technology—Japan. She is a lecturer at the School of Electrical and Informatics Engineering, Bandung Institute of Technology (ITB), Indonesia. She is also a co-founder at Prosa AI and a researcher at the Artificial Intelligence Center - ITB. Her research interest includes speech and natural language processing for Bahasa Indonesia.