Journal of
Information Systems Engineering
and Business Intelligence

Vol.10, No.3, October 2024
Available online at: http://e-journal.unair.ac.id/index.php/JISEBI

Unveiling User Sentiment: Aspect-Based Analysis and Topic Modeling of Ride-Hailing and Google Play App Reviews

Viktor Handrianus Pranatawijaya [1)], Nova Noor Kamala Sari [2)*], Resha Ananda Rahman [3)], Efrans Christian [4)], Septian Geges [5)]

[1)2)3)4)5)] *Informatics Engineering Department, Faculty of Engineering, Universitas Palangka Raya, Palangka Raya, Indonesia*
[1)]viktorhp@it.upr.ac.id, [2)]novanoorks@it.upr.ac.id, [3)] reshananda@mhs.eng.upr.ac.id, [4)] efrans@it.upr.ac.id, [5)] septian.geges@it.upr.ac.id

Abstract

Background: Mobile app usage is increasing in the digital age, with Ride-Hailing app becoming the primary example of this trend. To obtain valuable understanding of how people perceive and interact with mobile app, user reviews on platforms such as Google Play are usually analyzed. This analysis can assist developers to identify areas for improvement in both Ride-hailing and Google Play App. A promising method that can be used to analyze user perception in this instance is Aspect-Based Sentiment Analysis (ABSA).
Objective: This research aimed to apply ABSA to user reviews using Bidirectional Encoder Representations from Transformers (BERT) models. In this context, aspect identification and topic modeling were performed by using Latent Dirichlet Allocation (LDA). The model extracted topics from the reviews and used Generative Artificial Intelligence (GenAI) to define the aspects of the topics to further enhance the analysis. For consistency and accuracy, the method included sentiment annotation by a human annotator.
Methods: A total of two datasets were used in this research, with the first collected by scraping user reviews of Ride-Hailing App while the second was obtained from Kaggle, and to identify relevant topics, modeling was performed using LDA. These topics were then categorized into aspects using GenAI, covering areas, such as customer experience, service, payment, app features, task management, and event management. Subsequently, sentiment labeling was conducted using human annotators to provide a reliable baseline. BERT model was then used to classify sentiment with aspect hints, and the evaluation included calculations of accuracy, precision, recall, and F1-score.
Results: The results showed that BERT model achieved the highest accuracy of 97% in sentiment analysis across all datasets.
Conclusion: This research provided valuable understanding of user experience and established a strong ABSA framework for analyzing user reviews using LDA, Aspect Annotation, GenAI, and BERT sentiment models. Future research could expand this method to other app categories and incorporate real-time ABSA for continuous monitoring and dynamic feedback.

Keywords: User Reviews, Aspect-Based Sentiment Analysis (ABSA), Sentiment Analysis, Topic Modeling, Generative Artificial Intelligence (GenAI)

Article history: Received 1 April 2024, first decision 13 May 2024, accepted 23 August 2024, available online 28 October 2024

I. INTRODUCTION

Sentiment analysis is a strong method to examine the opinions and emotions of customer towards a specific product or service [1]. The analysis can uncover areas due to the large number of reviews on the Google Play Store. This data may be examined to better understand user sentiment and identify possible concerns. Different views from diverse user, including difficulties with the design and functions of online motorbike booking, can be recognized and assessed. Moreover, analyzing the data can provide feedback to a Ride-Hailing App that will help develop its app and increase user satisfaction. Sentiment analysis is an area of Natural Language Processing (NLP) that focuses on recognizing, extracting, and understanding the views, emotions, and attitudes presented in text [2]. This method is often used to detect polarity (positive, negative, or neutral) in texts [3] as well as to identify specific features or entities that are the focus of the emotion [4][5].

Aspect-Based Sentiment Analysis (ABSA) is focused on detecting and understanding sentiment toward certain features of an entity, such as products, services, or people [6][7]. Aspects are determined by defining the features [5][8] and by analyzing word clouds [9][10]. To understand user sentiment, ABSA method was adopted, and the method includes identifying aspects through topic modeling using Latent Dirichlet Allocation (LDA), extracting

* Corresponding author

ISSN 2443-2555 (online) 2598-6333 (print)
doi: http://dx.doi.org/10.20473/jisebi.10.3.328-339

topics, and evaluating the calculation of coherence score. The process determines the optimal number of topics to be explored [10][11]. Following the discussion, another research also discussed that the extracted topic keywords are then mapped to various aspects of an entity to perform ABSA on product reviews [12].

Generative Artificial Intelligence (GenAI) is an artificial intelligence technology that is capable of creating new data [13]. This new data can be text, images, video, audio, or 3D models. GenAI library is used to synthesize the text [14], which is applied as an aspect and distinguishes it from previous research. Additionally, the method extracts aspects from existing topic keywords by using GenAI library. Gemini is an example of GenAI and will be used in this research [15].

The next step is to annotate the sentiments associated with each aspect identified with human annotations. Consequently, ABSA is proposed, which is designed to generate perceptions from user reviews to provide a deeper understanding of user feedback. This process distinguishes between positive as well as negative sentiments and also identifies valuable aspects and areas that need improvement. Therefore, the research aimed to improve ABSA method for user review analysis. When developing ABSA, two experiments are conducted, the first is ABSA from user reviews of the Ride-Hailing App, and the second is the analysis from user reviews of the application store. In previous research, manual effort is required when aspects from vendor websites are collected and then topics are assigned to specific aspects [12]. As a result, automated aspect extraction and algorithmic methods to map topics to the extracted aspects are required.

After interpreting, the data is processed first when the number of positive and negative sentiments is unbalanced. Synthetic Minority Oversampling Technique (SMOTE) has inspired and continues to inspire the development of many oversampling methods and is the most cited method for solving the class imbalance problem [16][17]. Moreover, Bidirectional Encoder Representations from Transformers (BERT) models are used for sentiment and aspect classification. ABSA includes classifying sentiments associated with particular aspects of a text [18]. In general, understanding the context and meaning of words in natural language text is critical to BERT transformer design [19][20]. BERT has been used to perform the following tasks and has been performed appropriately, especially for the research [4][21].

In the experimental setup for each dataset, the data for topic modeling is first prepared by converting the preprocessed text into a Bag of Words (BoW) representation. Following this process, LDA is applied to extract an optimal number of topics using coherence values, and reviewing the keywords of topics with GenAI to obtain aspects. Subsequently, sentiment is annotated using human annotator, and as for sentiment classification with hints from aspects, BERT model is trained and evaluated. The data is then identified based on the evaluation results by calculating accuracy, precision, recall, and F1-score.

Based on the previous explanation, this research aimed to improve ABSA method for the analysis of user reviews by incorporating two important advances. First, BERT is used as a powerful contextual language model to perform sentiment analysis classification with hints from aspects in ABSA framework. This approach goes beyond lexicon-based methods by considering the context of words in user reviews, leading to more accurate sentiment identification for each aspect. Second, the potential problem of unbalanced sentiment distribution in user reviews is addressed using SMOTE. Moreover, this research aimed to achieve a deeper understanding of user reviews by distinguishing sentiment and also identifying valuable aspects and areas in need of improvement. Sentiment classification with hints from aspect is identified based on evaluation results by calculating accuracy, precision, recall, and F1-score.

II. Literature Review

A. Aspect-Based Sentiment Analysis (ABSA)

ABSA performs the tasks of effectively classifying sentiments on each aspect of a sentence [22]. The analysis is a field of NLP that includes a wide variety of operations. Additionally, the operations in ABSA are typically Opinion Target Extraction (OTE), Aspect Category Detection (ACD), and Sentiment Polarity (SP). OTE handles the extraction of aspect terms (e.g. entities or attributes), ACD identifies entities and the associated attributes, and SP clarifies aspect polarity [7].

Opinion target is the aspect/feature of a product in online customer reviews [23]. Aspect category defines a unique aspect of an entity and the aim of sentiment analysis [24]. In addition, LDA model is used for sentence aspect identification [25]. The topic keywords extracted from LDA are mapped to various entity aspects to perform aspect-specific sentiment analysis on the reviews [12]. Relating to this discussion, ABSA aims to identify and analyze sentiment toward certain aspects or features of an aimed entity in a certain text [26].

B. Related Works

In previous research on Aspect Category Classification with a Machine Learning method using an Indonesian language data set which was analyzed and tested [5]. Results from this examination showed that Random Forest outperformed other machine learning classification methods across domains. Moreover, the problem identified in this research is dataset imbalance which affects the results of accuracy, recall, and F1-score.

The research, Comparison of Naïve Bayes, SVM, and K-NN for ABSA of Gadgets [27] compared three machine learning algorithms for analyzing sentiment in YouTube comments on the Samsung Galaxy Z Flip 3 gadget. Following this exploration, people were more enthusiastic about the design of gadgets and more critical about the pricing, features, and brand image. SVM classification model outperformed Naive Bayes and K-NN models in terms of accuracy.

In this research, emotion detection based on aspects and correlations in tweets was used to understand public sentiment towards COVID-19 pandemic [9]. The fusion of aspect-based and correlation variables led to the highest accuracy rate of 96%. In addition, the document provided a valuable understanding of the perspective of public on COVID-19. A recommendation was made to investigate alternative machine learning and deep learning frameworks to identify sentiments and emotions based on aspects. Moreover, strategies such as transfer learning and domain adaptation should be explored to improve the performance of classification models in specific domains. To achieve a comprehensive understanding of public sentiment, the integration of ABSA and emotion detection with other natural language processing tasks such as topic modeling and opinion mining was recommended. Accuracy dropped to 96% in scenarios where one component was missing. Additionally, the research uncovered major factors influencing public sentiment and emotional state concerning COVID-19, such as fear, anger, hopelessness, and anxiety.

IndoBERT model was used in the research project [4] to conduct ABSA on internet evaluations of Bromo Tengger Semeru National Park (BTSNP). This research aimed to discover the factors that influenced the attitudes of tourists concerning BTSNP and to determine total satisfaction with the park. IndoBERT model was 91.48% accurate in identifying the sentiment of aspects in BTSNP evaluations. Relating to this discussion, the most often addressed topics were attractions, amenities, and pricing.

III. METHODS

Several steps in this method included data collection, preprocessing, aspect identification, topic modeling, sentiment annotation, as well as sentiment and aspect classification. Fig. 1 showed a detailed explanation to help understand how the process worked.

A. Data Collecting

Data collection was applied to Ride-Hailing and Google Play App Reviews. The first step was to collect data on Ride-hailing App (Gojek) through the Google Play Store for the first dataset. In the context of this research, data collection was conducted using scraping [28][29]. The language used in user reviews was English, due to the many advantages of using it. These advantages included increasing global visibility, standardizing the user experience across markets, and ease of analysis, as more advanced analysis tools and technologies were available. The data extraction method was performed using a Python module called google-play-scraper. In this research, the review data was collected from January 2020 to December 2023. During the data collection phase, the dataset was saved in CSV format. A total of 14,147 reviews were successfully collected from the collection procedure.

The scraped data collected was used as the basis for this research. Gojek reviewed data comprising the username review, text, rating, and publishing time. Among these attributes, the review text attribute was used in the research as shown in Table 1.

TABLE 1
GOJEK DATA SCRAPING SAMPLE

Username	Review Text	Review Rating	Review Publishing Time
Obet Nyawo	It keeps crashing when I'm about to order. Seems as when there's a bug. Please fix it asap because it's annoying	2	12/12/2023 4:15
Rahma omceee	After updating, yesterday I still can use it, now I can't use the apps, when I open just blank and then back to the phone menu.	1	12/12/2023 2:47
Any	The app crashed and caused my order to be placed twice. My GoPay was charged 2 times, yet I did not enter my PIN. This sucks! Fix the app	1	12/14/2023 5:00
Priam Sampaio	The new mode of fast-booking is abusive - I don't even have the time to hit "return" when Gojek booked me a ride and refuses to show me the price until the trip is finished.	1	12/12/2023 0:56
Avior Jonas	Not good apps, the gocar driver that I booked was occasionally far away from my position, and we need is good response for each driver.	3	12/21/2023 23:44

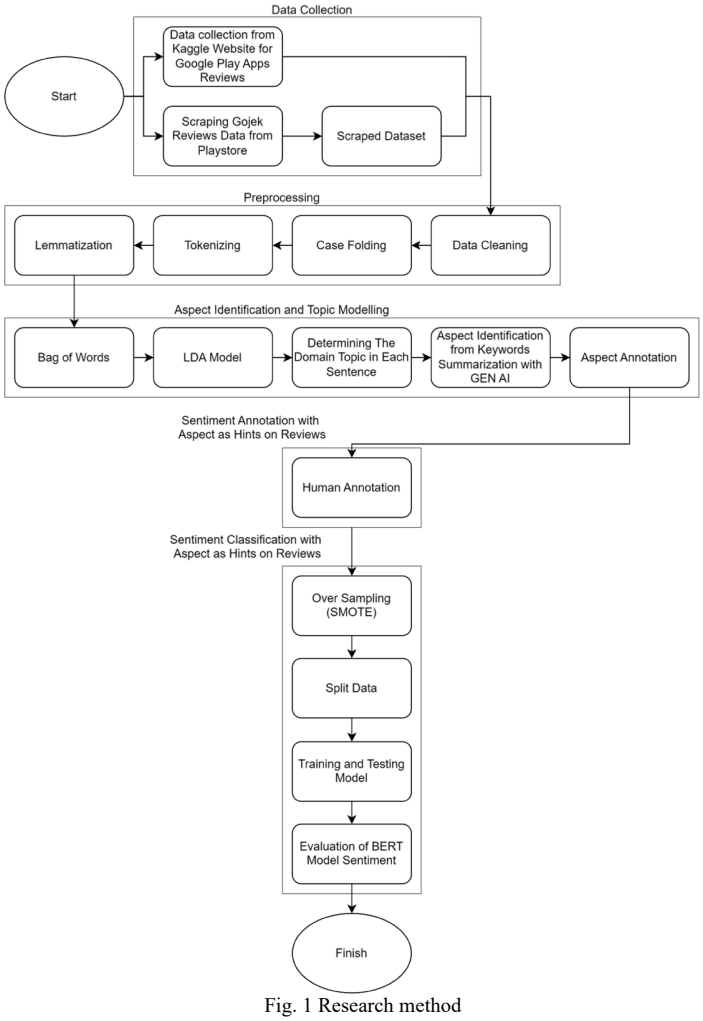


Fig. 1 Research method

The second collection of data was data that was labeled and obtained from Kaggle website. The dataset consisted of 16.390 entries, which were used for comparison. In this research, a few samples from the "Google Play Apps Reviews" dataset were shown [30]. This research used the review text attribute as shown in Table 2.

TABLE 2
GOOGLE PLAY APPS REVIEWS DATA SAMPLE

Username	Review Text	Review Rating	Review Publishing Time
Daniel Gonsoulin	Too complicated. Had one day, three day, and week view, no monthly view.	1	11/18/2020 18:04
moss moss	Impossible to cancel. Beware before you take 7-day free option giving your credit card details	1	11/17/2020 9:31
Charles Bartusch	I loved this app however now it doesn't connect to Alexa lists anymore	1	12/12/2020 15:14
Steff Cee	Won’t add people to tasks!!! Wanted to delete a sub-task and it deleted the entire to-do list. ANNOYING	1	12/23/2020 6:39
Avinash soni	Too expensive and even not able to delete the account after setting the password it says the wrong password	1	11/27/2020 4:48

B. Data Preprocessing

Preprocessing was aimed at extracting data before processing it further. Data cleaning, case folding, tokenizing, and lemmatization were all part of the preprocessing [31]. Moreover, data cleaning aimed to clean up text by performing various cleaning processes, such as cleaning up emojis, hashtags, mentions, URLs, and characters. Letters and spaces, dates, numbers, date/time, and additional punctuation were not used in the sentiment analysis process. The

library was used in the data-cleaning process to work with regular expressions, which were powerful tools for matching text patterns.

Case Folding was used for converting text into a more compact form. This particular process used a string as input and altered it entirely into a condensed string, using built-in lower of Python. Following the process, tokenizing which used Spacy library, included dividing the text into smaller parts, typically consisting of words or specific symbols, creating tokens that were singular words. Lemmatization using Spacy library served the purpose of transforming words into a fundamental form that corresponded to word type. The outcomes of the data preprocessing process were shown in Table 3.

TABLE 3
SAMPLE DATA PREPROCESSING

Preprocessing	Input	Output
Cleaning	The new mode does fast-booking is abusive - I don't even have the time to hit "return" that gojek booked me a ride and refuses to show me the price until the trip is finished.	The new mode does fastbooking is abusive I don't even have the time to hit return when Gojek booked me a ride and refused to show me the price until the trip was finished
Case Folding	The new mode does fastbooking is abusive I didn't even have the time to hit return when Gojek booked me a ride and refused to show me the price until the trip was finished	the new mode does fastbooking is abusive I didn't even have the time to hit return when Gojek booked me a ride and refused to show me the price until the trip was finished
Tokenizing	the new mode does fastbooking is abusive I don't even have the time to hit return when Gojek booked me a ride and refused to show me the price until the trip was finished	'the', 'new', 'mode', 'does', 'fastbooking', 'is', 'abusive', ' ', 'i', 'do', 'nt', 'even', 'have', 'the', 'time', 'to', 'hit', 'return', 'that', 'gojek', 'booked', 'me', 'a', 'ride', 'and', 'refuses', 'to', 'show', 'me', 'the', 'price', 'until', 'the', 'trip', 'is', 'finished']
Lemmatization	['the', 'new', 'mode', 'does', 'fastbooking', 'is', 'abusive', ' ', 'i', 'do', 'nt', 'even', 'have', 'the', 'time', 'to', 'hit', 'return', 'that', 'gojek', 'booked', 'me', 'a', 'ride', 'and', 'refuses', 'to', 'show', 'me', 'the', 'price', 'until', 'the', 'trip', 'is', 'finished']	The new mode does fastbook be abusive i did not even have the time to hit return when Gojek booked me a ride and refused to show me the price until the trip was finished

C. Aspect Identification and Topic Modeling

During the aspect identification and topic modeling stages, LDA algorithm was used as a method to group topics derived from two preprocessed data sets. In addition to lexical resources, the input consisted of a vector representation of the lexicon. The results of the word weighting were used by the topic modeling framework for data mapping and analysis. Additionally, the weighting method was based on the BoW method [21], as implemented in the Gensim library.

This segment was dedicated to BoW methodology using resolved feedback data. Following this process, the data was then transformed into a lexical resource. After creating lexicon, the subsequent stage included generating a BoW or term frequency from the corpus. The output of this procedure was the data stored in the variables id2word and Bow corpus.

LDA was applied to extract topics present in the reviews and then model evaluation was conducted through the examination of coherence scores to evaluate its efficacy. As part of this LDA model, various essential parameters were incorporated, including BoW, number of topics, dictionary, chunk size, passes, random state, and alpha. The parameter "passes" was used to determine the number of iterations or training sessions the model passed through. Concerning parameters related to the count of topics, LDA model functioned to detect patterns of topics based on the distribution of words in the document, thereby striving to identify topics following the specified number of topics. During this procedure, coherence scores played a critical role in evaluating the interpretability and coherence of the topics discovered by the model.

The model derived results from the aforementioned process including the use of a limit of five, which served to limit the maximum number of topics to be explored. Additionally, an initial value of two was used to specify the minimum number of topics to be researched, with a step size of one used to increase the number of topics. The outcomes of coherence value relating to each topic were shown in Table 4.

The result of implementing LDA model was determined by considering the highest coherence value obtained from LDA modeling, namely 3 topics in each dataset. As a result, the conclusion drawn from the proposed review dataset was that the 3 topics appeared most frequently in each dataset.

After obtaining the number of topics by considering the highest coherence value, the keywords that appeared frequently in each topic were identified. These keywords were used to determine relevant aspects of each topic. Moreover, aspect identification was conducted by reviewing topic keywords. GenAI was used to review the keywords in each topic derived from LDA, and the model applied was GenAI library of Gemini-pro. After obtaining the aspects of GenAI, another validation was performed using human judgment to determine whether the aspects were appropriate and feasible to use. In this research, the prompt consisted of a series of keywords per topic which included " Review

the topics for finding the aspect, give me two words". Following this process, the determination of aspects was performed on each dataset. More understanding was shown in Algorithm 1 for aspects generated by GenAI.

TABLE 4
THE OUTCOMES OF COHERENCE VALUE RELATING TO EACH TOPIC

Dataset	Num Topics	Coherence Value
Gojek App Reviews	2	0.488
	3	0.533
	4	0.521
Google Play Apps Reviews	2	0.508
	3	0.529
	4	0.469

Algorithm 1
Aspects Generated by GenAI

Setup: *Import libraries, configure API key.*
Find Relevant Models: *Identify GenAI models that can generate content.*
Define Text Formatting: *Create a function to format text for show.*
Load Gemini Model: *Create a Generative Model object for Gemini.*
Loop through Topics:
For each topic*:*
Construct a prompt for Gemini combining the topic and instructions for aspect keywords.
Generate content using generate_content with the prompt.
Format and show the generated text (keywords) for each topic.

From the previous report, aspects were obtained using GenAI from keywords gotten from topics applying LDA, as each dataset produced exceptional aspects. Additionally, the aspects from the first dataset included customer experience, service, and payment. App features, task management, and event management were the aspects of the second dataset. More explanation was discussed in the research as shown in Table 5.

TABLE 5
THE DISTRIBUTION OF WORDS BASED ON THE NUMBER OF TOPICS CAN BE EXAMINED

Dataset	Topic	Aspect (Generated by GenAI)	Keywords Per Topic (Generates by LDA)
Gojek App Reviews	1	Customer Experience	driver, get, order, bad, time, customer, cancel, star, price, fee, find, just, give, go, even, ride, take, too, location, see
	2	Service	good, app, service, very, great, use, gojek, easy, so, help, helpful, thank, make, life, application, helpful, love, always, nice, now
	3	Payment	gopay, top, change, pay, account, email, try, say, money, payment, delete, later, back, phone, unable, sign, reason, already, transaction, history
Google Play Apps Reviews	1	App Features	app, use, good, great, very, planner, version, phone, need, well, get, just, ad, pay, feature, really, like, more, make, keep
	2	Task management	task, time, widget, add, show, update, list, go, change, see, option, also, day, now, fix, new, able, notification, screen, date
	3	Event Management	calendar, work, event, love, so, sync, easy, day, try, note, ve, m, view, reminder, set, far, still, month, week, find

D. Sentiment Annotation

Sentiment annotation was conducted using two human annotators and kappa [32], [33], [34] to measure agreement [34]. The kappa score calculation was performed by importing cohen_kappa_score function from sklearn.metrics library in Python. The interpretation of kappa values was explained appropriately in Table 6.

When determining the sentiment of reviews, the aspects of the review were also considered allowing the sentiment determined to correlate with the aspects. The annotation results of Gojek App Reviews dataset included 7945 reviews with negative sentiment and 6170 with positive sentiment. In addition, kappa value of the first dataset was 0.948054205018953, and the strength of agreement was almost perfect. On the Google Play App Reviews dataset, 9860 reviews had negative sentiment and 6346 reviews had positive sentiment. Following this discussion, Table 7 showed some results from ABSA. Kappa value of the first dataset was 0.9926120845062325 and strength of agreement was almost perfect.

TABLE 6
INTERPRETATION OF KAPPA VALUES

Kappa Value			Strength of Agreement
< 0.00			Poor
0.00	-	0.20	Slight
0.21	-	0.40	Fair
0.41	-	0.60	Moderate
0.61	-	0.80	Substantial
> 0.81	-	0.99	Almost perfect
1.00			Perfect

TABLE 7
ASPECT-BASED SENTIMENT ANALYSIS (ABSA)

Dataset	Review	Aspect	Sentiment
Gojek App Reviews	On-Time pick-up & Driver hospitality & secure_safe.	Customer Experience	Positive
	Thank you for saving my life, my old number has already been deactivated, but because the "Send to Whatsapp" feature, I was able to get back to my account safely again.	Service	Positive
	Usually fine use this app, but now cannot even open cause it keeps crashing on my phone.	Payment	Negative
	The features and functionality are pretty good, with support for Google Tasks. But the home screen widget doesn't load at all on Android 9 after restarting the device. One has to remove it and reinstall it, which is a pain. And the support team was not helpful at all. The service went from recommending a solution that the service didn't understand the app to being completely unresponsive.	App Features	Negative
Google Play App Reviews	I used to love ticktick. I have it on mobile Android and Windows 10 client. The systems used to sync beautifully. Around 23 Aug, the desktop client and mobile stopped syncing and acted as two separate accounts. In desperation, I reinstalled the desktop client. I couldn't log into client or through browser. It complained of network configuration and proxies. I don't use proxies. I can log into any account online except ticktick. The help rep is unresponsive. Super disappointed. I'm on premium. I want refund.	Task management	Negative
	Professional-looking business calendar, love how you can add export, and import your data, pretty good and useful, and more options than your standard calendar on your phone.	Event Management	Positive

E. Sentiment Classification with Aspects as Hints on Reviews

Sentiment classification with aspects as hints on reviews consisted of model training and model evaluation used to classify sentiment. Hints on reviews were conducted by combining aspects and reviews into one feature. Therefore, in a review, the aspects of the review were also considered allowing the sentiment to be determined according to these aspects. The models that used BERT classifier were trained by splitting into separate processes per model and evaluated using evaluation metrics for instance accuracy, precision, recall, and F1-score. These processes included datasets such as Google Play App Reviews and Gojek App Reviews, which contained aspect and sentiment labels. Relating to the process, the datasets were then divided into training and test data. The data splitting was conducted by dividing the training data to 85% and the test data to 15%. In this research, the features included combined text of aspect and review content. At the time of classification, data checking was performed, and the aim was sentiment label which was positive as well as negative. The process occurred because the training data needed to be balanced, and a way to achieve this was by using SMOTE to balance it.

1) Synthetic Minority Oversampling Technique (SMOTE)

The class distribution in the first dataset was 7945 negatives and 6170 positives. In the second dataset, 9860 were negatives and 6346 were positives. After splitting the data, the sentiment class distribution in the training data was checked. Moreover, when the class distribution was not balanced, then oversampling was performed.

The training data in the first dataset was 6751 negatives and 5246 positives. Then in the second dataset, there were 8346 negatives and 5429 for training data. Based on the class distribution, a conclusion was made that the classes were not balanced. Therefore, data balancing was conducted by importing SMOTE class in Python and using the imblearn.over_sampling library.

To balance the classes in the first dataset, SMOTE added synthetic samples to the positive class allowing the total to be close or equal to the negative class (6751). This process implied that SMOTE generated about 6751 - 5246 = 1325 new synthetic samples for the positive class. The method was applied to the second dataset to oversample the minority class (positive class). The process aimed to make the amount of data in both groups (positive and negative) almost equal, and about 2917 new data was added in the positive group. The class distribution results for positive and negative sentiments were 6571 in the first dataset and 8346 in the second dataset. Concerning this process, more details were shown in Table 8.

TABLE 8
COMPARISON OF CLASS DISTRIBUTION TRAINING DATA BEFORE AND AFTER SMOTE IMPLEMENTATION

Dataset	Class Distribution	Before SMOTE	After SMOTE
Gojek App Reviews	Positive	5246	6571
	Negative	6571	6571
Google Play App Reviews	Positive	5429	8346
	Negative	8346	8346

2) Bidirectional Encoder Representations from Transformers (BERT)

BERT model used in this research was distilbert-base-uncased-finetuned-sst-2-english [35]. In the BERT framework, a tokenizer entity was generated to divide the text into smaller tokens, which were integrated into BERT model for various analyses such as text classification and generation. Additionally, the text input was processed according to the training methodology of the model by using a tokenizer modified to BERT model. Major BERT parameters used included learning rate (1e-5), batch size of 32, maximum sequence length of 128, and 3 training epochs.

F. Evaluation

During the final phase, an assessment was conducted to identify the optimal sentiment analysis model. Numerous popular performance measures were developed for multi-class classification [36]. Moreover, the evaluation criteria used by the authors included accuracy, precision, recall, and F1-score to measure the effectiveness of the sentiment analysis model developed in this research.

The classification report function in scikit-learn was used to produce a detailed performance evaluation. This report provided a review of various evaluation metrics for the classification model. The function was particularly useful for assessing model performance in multi-class or binary classification. In addition, the function provided a detailed overview of the performance of the model in terms of precision, recall, and F1-score. This function also provided support for each class as well as macro and weighted averages, which was very useful for evaluating the classification model as a whole.

The model worked by calculating the precision, recall, and F1-Score for each class (positive and negative) in the test data. After obtaining the value, the accuracy and support were calculated during this process. The next step was to calculate the macro and weighted average values. In this research, macro average was a simple arithmetic average calculation of performance metrics (precision, recall, and F1-score) for each class independently, without considering different class weights. Different from the macro average which achieved equal weight to each class, the weighted average considered the proportion of classes in the dataset, consequently giving a more representative picture of the performance of the model to the actual class distribution.

IV. RESULTS

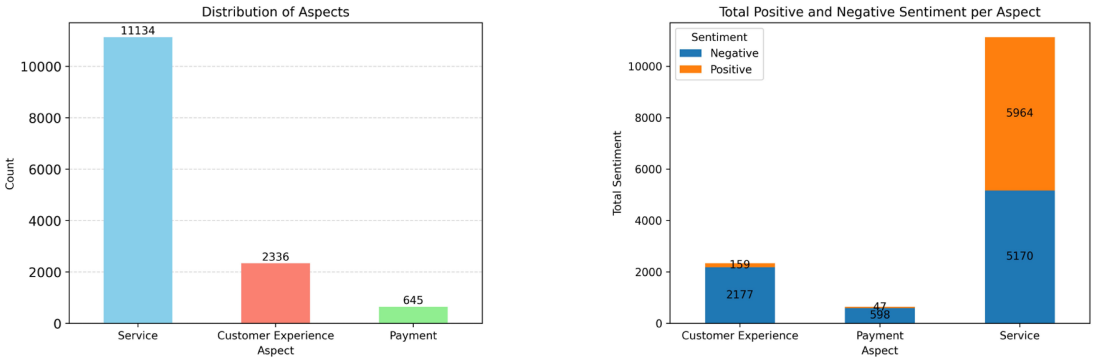
The distribution of the number of reviews for each aspect was shown in the following Table. The aspect distribution based on Gojek App Reviews consisted of 11134 services, 2336 customer experience, and 645 payments. In addition, the sentiment distribution per aspect included 159 positives and 2177 negatives sentiment for customer experience, 5964 positives and 5170 negatives sentiment for service, and 47 positives and 598 negatives sentiment for payment. Table 9 showed the distribution of sentiments per aspect and the graphical presentation as also shown in Fig. 2.

The distribution of the number of reviews for each aspect was shown in the following Table. The aspect distribution based on Google Play App Reviews on the aspect distribution consisted of 11757 app features, 3785 task management, and 664 event management. Additionally, the sentiment distribution per aspect consisted of 5373 positives and 6384 negatives sentiment for app features, 762 positives and 3023 negatives sentiment for task management, and 211 positives and 453 negatives sentiment for event management. Table 10 showed the distribution of sentiments per aspect also as shown in Fig. 3.

The evaluation of BERT model used to classify sentiment with aspects as hints on reviews per labeled dataset was performed. In the sentiment evaluation, the features of BERT model were combined text that consisted of aspect as well as review text, and the aim was the sentiment labels. Moreover, macro average calculation was performed to calculate average precision, recall, and F1-score across classes without considering class proportions. The weighted average calculation was conducted to find average precision, recall, and F1-score across classes that considered class proportions. Following this discussion, support showed how many samples in the dataset belong to that class. This result provided a more accurate, general picture when there was class imbalance. Table 11 showed a more complete explanation of how the process worked.

TABLE 9
DISTRIBUTION OF SENTIMENT PER ASPECT FROM GOJEK APP REVIEWS

Aspect	Positive	Negative	Ratio (Positive:Negative)	Dominant Sentiment
Customer Experience	159	2177	1 : 13.7 (approx.)	Negative
Service	5964	5170	1.2 : 1 (approx.)	Slightly Positive
Payment	47	598	1 : 12.7 (approx.)	Negative

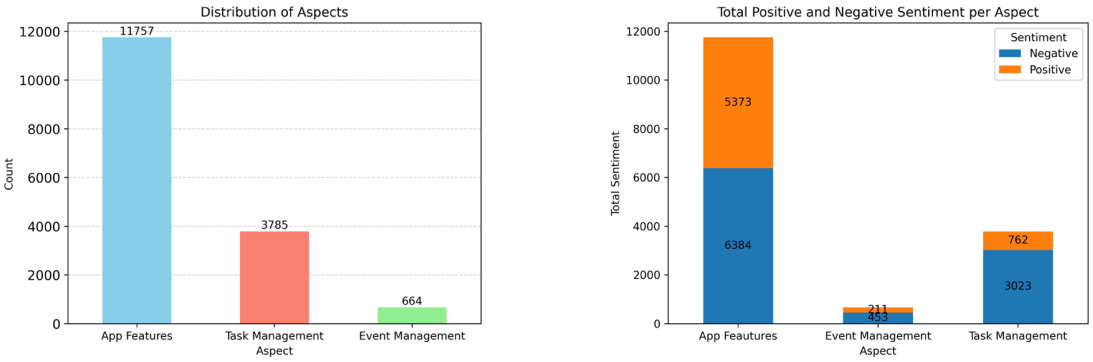


(a) (b)

Fig. 2 Gojek app reviews on (a) The aspect distribution; (b) Total positive and negative sentiment per aspects

TABLE 10
DISTRIBUTION OF SENTIMENT PER ASPECT FROM GOOGLE APP REVIEWS

Aspect	Positive	Negative	Ratio (Positive:Negative)	Dominant Sentiment
App Features	5373	6384	0.9 : 1 (approx.)	Negative
Task Management	762	3023	1 : 4 (approx.)	Negative
Event Management	211	453	0.5 : 1 (approx.)	Negative



(a) (b)

Fig. 3 Google Play app reviews on (a) The aspect distribution; (b) Total positive and negative sentiment per aspects

TABLE 11
BERT MODEL SENTIMENT EVALUATION

Dataset	Accuracy	Class	Precision	Recall	F1-Score	Support
		Negative	99 %	95 %	97 %	1194
		Positive	94 %	99 %	96 %	924
Gojek App Reviews	97 %					2118
		Macro Avg	97 %	97 %	97 %	2118
		Weighted Avg	97 %	97 %	97 %	2118
		Negative	99 %	97 %	98 %	1514
		Positive	95 %	98 %	96 %	917
Google Play App Reviews	97 %					2431
		Macro Avg	97 %	97 %	97 %	2431
		Weighted Avg	97 %	97 %	97 %	2431

V. DISCUSSION

Several ways to identify aspects included aspect identification, which was determined by analyzing wordclouds [9][10], and also determined through topic modeling [11][12] which were mapped into areas based on the specified aspects. After the aspect was obtained during the process, it explored the sentiment in ABSA. In this research, aspects were obtained from reviews derived from LDA and GenAI and not determined by the explorers or word clouds.

Based on sentiment analysis of Gojek App Review dataset and Google Play App Review dataset using BERT model, the first dataset showed the best performance, with accuracy, precision, recall, F1-score value was 97%. The use of SMOTE by considering the class proportion led to the best performance. In addition, the evaluation metrics for the first dataset generally outperformed the second dataset, and the accuracy achieved exceeded previous research on SVM [5][27] and the results on BERT [4][9]. The method used ABSA and showed accuracy ranging from 74% to 81% for each aspect [12] and the proposed method outperformed with 97% accuracy.

For Gojek App Reviews, customer experience had the largest disparity between positive and negative sentiment, showing a strongly negative experience for customer. The service aspect, while slightly positive, still showed a balanced distribution, proposing room for improvement despite more positive experiences compared to negative factors. Similarly, payment showed a dominant negative sentiment, implying significant issues with the payment process. The aspects ranked from worst to best were as follows, customer experience, payment, and service.

All three aspects showed a negative sentiment distribution, implying total dissatisfaction with the app in these areas for the Google Play App. Task management showed the most significant disparity, with a ratio of 1:4, implying a strong need for improvement in this feature. Additionally, app features, although not as extreme, still had a negative sentiment ratio of 0.9:1, signifying that the user was encountering limitations or issues. Event management, while also negative, had the closest ratio to a balanced distribution at 0.5:1, showing that user found these features helpful, but there remained room for improvement. The aspects were ranked from worst to best as follows, task management, app features, and event management.

VI. CONCLUSIONS

In conclusion, LDA and GenAI were used to annotate aspects based on the research results. This process was conducted and produced three aspects on each dataset run by GenAI according to the three clusters obtained from the use of LDA where the coherent value was used to determine the number of topics. Moreover, sentiment annotation was performed using BERT based on the highest metric evaluation results obtained. The best performing BERT model distilbert-base-uncased-finetuned-sst-2-english was used for aspect and sentiment classification with the parameters of learning rate (1e-5), batch size 32, maximum sequence length 128, and 3 training epochs. Therefore, this research provided a strong framework for ABSA on user reviews in different domains.

The deeper understanding achieved from the user reviews of the two datasets in this research was the identification of the valuable aspects and those that need to be improved. In addition, the aspects were influenced by the sentiment of the user. This process was a valuable perception for Ride-hailing and Google Play App.

Based on the sentiment distribution on Ride-Hailing App, customer experience had the most significant negative bias, payment showed strong negative sentiment, and service, although slightly positive, still needed improvement. Recommendations included focusing on improving customer experience by investigating the root causes of dissatisfaction, addressing issues with the payment process to reduce customer frustration, and maintaining service quality while striving for further improvement.

According to the sentiment distribution on Google Play App, task management had the most significant negative bias, app features had strong negative sentiment, and event management showed the least negative sentiment, though it was still negative generally. For improvement, addressing issues in task management should be prioritized, analyzing as well as improving app features based on user feedback, and finally evaluating event management for areas needing development despite its relatively better sentiment.

Future research could be extended to continuous monitoring by implementing ABSA in real time to continuously measure user satisfaction and respond quickly to problems. Another area would be to prioritize improvements based on user feedback, focusing on frequently expressed negative aspects.

Author Contributions: *Viktor Handrianus Pranatawijaya*: Conceptualization, Methodology, Writing - Original Draft, Writing - Review & Editing, Supervision. *Nova Noor Kamala Sari*: Software, Investigation, Data Curation, Writing - Review & Editing, Writing - Original Draft. *Resha Ananda Rahman:* Software, Writing - Original Draft. *Efrans Christian*: Writing - Review & Editing, Investigation, Data Curation. *Septian Geges*: Investigation, Data Curation.

All authors have read and agreed to the published version of the manuscript.

Funding: This research received no specific grant from any funding agency.

Conflicts of Interest: The authors declare no conflict of interest.

Data Availability: Dataset used in this study can be accessed at https://www.kaggle.com/datasets/beliau/gojek-app-reviews and https://www.kaggle.com/datasets/therealsampat/google-play-apps-reviews.

Institutional Review Board Statement: Not applicable.

Informed Consent: There were no human subjects.

Animal Subjects: There were no animal subjects.

ORCID:
Viktor Handrianus Pranatawijaya: https://orcid.org/0000-0002-3301-0702
Nova Noor Kamala Sari: https://orcid.org/0009-0006-0851-3216
Resha Ananda Rahman: -
Efrans Christian: https://orcid.org/0000-0001-7781-4756
Septian Geges: https://orcid.org/0009-0003-4636-3931

REFERENCES

[1] R. Bose, R. K. Dey, S. Roy, and D. Sarddar, "Sentiment analysis on online product reviews," in Information and Communication Technology for Sustainable Development, 2020, pp. 559–569. doi: 10.1007/978-981-13-7166-0_56.

[2] A. Hermawan, I. Jowensen, J. Junaedi, and Edy, "Implementasi text-mining untuk analisis sentimen pada twitter dengan algoritma Support Vector Machine," *Jurnal Sains dan Teknologi*, vol. 12, no. 1, pp. 129–137, Apr. 2023, doi: 10.23887/jstundiksha.v12i1.52358.

[3] P. H. Prastyo, A. S. Sumi, A. W. Dian, and A. E. Permanasari, "Tweets responding to the Indonesian government's handling of Covid-19: sentiment analysis using SVM with normalized poly kernel," *Journal of Information Systems Engineering and Business Intelligence*, vol. 6, no. 2, p. 112, Oct. 2020, doi: 10.20473/jisebi.6.2.112-122.

[4] C. A. Bahri and L. H. Suadaa, "Aspect-based sentiment analysis in Bromo Tengger Semeru National Park Indonesia based on google maps user reviews," *Indonesian Journal of Computing and Cybernetics Systems*, vol. 17, no. 1, p. 79, Feb. 2023, doi: 10.22146/ijccs.77354.

[5] S. A. P. Perdana, T. B. Aji, and R. Ferdiana, "Aspect category classification with machine learning approach using Indonesian language dataset," *Jurnal Nasional Teknik Elektro dan Teknologi Informasi*, vol. 10, no. 3, pp. 229-235, 2021, doi: 10.22146/jnteti.v10i3.1819.

[6] I. A. Kandhro, F. Ali, M. Uddin, A. Kehar, and S. Manickam, "Exploring aspect-based sentiment analysis: An in-depth review of current methods and prospects for advancement," Knowl. Inf. Syst., vol. 66, no. 7, pp. 3639–3669, Jul. 2024, doi: 10.1007/s10115-024-02104-8.

[7] R. Bharathi, R. Bhavani, and R. Priya, "Leveraging deep learning models for automated aspect based sentiment analysis and classification," *SSRG International Journal of Electrical and Electronics Engineering*, vol. 10, no. 5, pp. 120–130, May 2023, doi: 10.14445/23488379/IJEEE-V10I5P111.

[8] S. Cahyaningtyas, D. Hatta Fudholi, and A. Fathan Hidayatullah, "Deep learning for aspect-based sentiment analysis on Indonesian hotels reviews," *Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control*, Aug. 2021, doi: 10.22219/kinetik.v6i3.1300.

[9] Salsabila, S. M. P. Tyas, Y. Romadhona, and D. Purwitasari, "Aspect-based sentiment and correlation-based emotion detection on tweets for understanding public opinion of Covid-19.," *Journal of Information Systems Engineering and Business Intelligence*, vol. 9, no. 1, pp. 84–94, 2023, doi: 10.20473/jisebi.9.1.84-94.
[10] W. A. Nawang Sari and H. Dwi Purnomo, "Topic modeling using the latent dirichlet allocation method on Wikipedia pandemic Covid-19 data in Indonesia," *Jurnal Teknik Informatika*, vol. 3, no. 5, pp. 1223–1230, Oct. 2022, doi: 10.20884/1.jutif.2022.3.5.321.
[11] F. Rashif, G. Ihza Perwira Nirvana, M. Alif Noor, and N. Aini Rakhmawati, "Implementasi LDA untuk pengelompokan topik cuitan akun bot Twitter bertagar #Covid-19," *Cogito Smart Journal*, vol. 7, no. 1, 2021.
[12] V. S. Anoop and S. Asharaf, "Aspect-oriented sentiment analysis: A topic modeling-powered approach," *Journal of Intelligent Systems*, vol. 29, no. 1, pp. 1166–1178, Jan. 2020, doi: 10.1515/jisys-2018-0299.
[13] R. Gozalo-Brizuela and E. C. Garrido-Merchán, "A survey of generative AI applications," Jun. 2023, [Online]. Available: http://arxiv.org/abs/2306.02781
[14] P. Denny *et al.*, "Computing education in the era of Generative AI," Jun. 2023, [Online]. Available: http://arxiv.org/abs/2306.02608
[15] Gemini Team *et al.*, "Gemini: A family of highly capable multimodal models," Dec. 2023, [Online]. Available: http://arxiv.org/abs/2312.11805
[16] F. Rodríguez-Torres, J. F. Martínez-Trinidad, and J. A. Carrasco-Ochoa, "An oversampling method for class imbalance problems on large datasets," *Applied Sciences*, vol. 12, no. 7, p. 3424, Apr. 2022, doi: 10.3390/app12073424.
[17] P. Gnip, L. Vokorokos, and P. Drotár, "Selective oversampling approach for strongly imbalanced data," *PeerJ Comput Sci*, vol. 7, pp. 1–22, 2021, doi: 10.7717/PEERJ-CS.604.
[18] J. Z. Maitama, N. Idris, A. Abdi, and A. T. Bimba, "Aspect extraction in sentiment analysis based on emotional affect using supervised approach," in *2021 4th International Conference on Artificial Intelligence and Big Data, ICAIBD 2021*, Institute of Electrical and Electronics Engineers Inc., May 2021, pp. 372–376. doi: 10.1109/ICAIBD51990.2021.9458996.
[19] O. J. Prasad, S. Nandi, V. Dogra, and D. S. Diwakar, "A systematic review of NLP methods for sentiment classification of online news articles," in *2023 14th International Conference on Computing Communication and Networking Technologies, ICCCNT 2023*, Institute of Electrical and Electronics Engineers Inc., 2023. doi: 10.1109/ICCCNT56998.2023.10308056.
[20] T. Sai Aparna, K. Simran, B. Premjith, and K. P. Soman, "Aspect-based sentiment analysis in Hindi: Comparison of machine/deep learning algorithms," 2021, pp. 81–91. doi: 10.1007/978-981-33-4305-4_7.
[21] M. P. Geetha and D. Karthika Renuka, "Improving the performance of aspect based sentiment analysis using fine-tuned BERT Base Uncased model," *International Journal of Intelligent Networks*, vol. 2, pp. 64–69, Jan. 2021, doi: 10.1016/j.ijin.2021.06.005.
[22] T. H. Khan and S. Ridhorkar, "Improving the efficiency of aspect-based sentiment analysis using ensemble deep learning," Int J Intell Syst Appl Eng, vol. 12, no. 10s, pp. 526–538, Jan. 2024.
[23] Q. Liu, Z. Gao, B. Liu, and Y. Zhang, "Automated rule selection for opinion target extraction," Knowledge-Based Systems, vol. 104, pp. 74–88, Jul. 2016, doi: 10.1016/j.knosys.2016.04.010.
[24] W. Zhang, X. Li, Y. Deng, L. Bing, and W. Lam, "A survey on aspect-based sentiment analysis: Tasks, methods, and challenges," Mar. 2022, [Online]. Available: http://arxiv.org/abs/2203.01054
[25] A. A. Bakar, LK. Soon, and HN. Goh, "An exploratory study on Latent-Dirichlet Allocation Models for aspect identification on short sentences," in *Computational Science and Tecnology: 4th ICCST 2017*, 2018, pp. 314–323. doi: 10.1007/978-981-10-8276-4_30.
[26] M. P. Abraham, K. R. Udaya Kumar Reddy, and S. Abraham, "Aspect based sentiment analysis system using machine learning techniques," in *2023 IEEE International Conference on Distributed Computing, VLSI, Electrical Circuits and Robotics (DISCOVER)*, IEEE, Oct. 2023, pp. 25–30. doi: 10.1109/DISCOVER58830.2023.10316716.
[27] J. W. Iskandar and Y. Nataliani, "Perbandingan Naïve Bayes, SVM, dan k-NN untuk analisis sentimen gadget berbasis aspek," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 5, no. 6, pp. 1120–1126, Dec. 2021, doi: 10.29207/resti.v5i6.3588.
[28] D. Deviacita *et al.*, "Implementasi web scraping untuk pengambilan data pada situs marketplace," *Jurnal Sistem dan Teknologi Informasi (JustIN)*, vol. 7, no. 4, 2019, doi: 10.26418/justin.v7i4.30930.
[29] A. Ariyo Munandar and C. Edi Widodo, "Sentimen analisis aplikasi belajar online menggunakan klasifikasi SVM," *Journal of Information Technology and Computer Science (JOINTECS)*, vol. 7, no. 2, pp. 77–84, 2023, doi: 10.31328/jointecs.v8i2.4747.
[30] S. Patel, "Google Play apps reviews," *Kaggle*, 2023, Accessed: Feb. 20, 2024. [Online]. Available: https://www.kaggle.com/datasets/therealsampat/google-play-apps-reviews
[31] L. Hermawan and M. B. Ismiati, "Pembelajaran text preprocessing berbasis simulator untuk mata kuliah information retrieval," *Jurnal Transformatika*, vol. 17, no. 2, pp. 188–199, 2020, doi: 10.26623/transformatika.v17i2.1705.
[32] G. Rau and Y. S. Shih, "Evaluation of Cohen's kappa and other measures of inter-rater agreement for genre analysis and other nominal data," *J Engl Acad Purp*, vol. 53, p. 101026, Sep. 2021, doi: 10.1016/j.jeap.2021.101026.
[33] Z. Yang and M. Zhou, "Kappa statistic for clustered matched-pair data," *Stat Med*, vol. 33, no. 15, pp. 2612–2633, Jul. 2014, doi: 10.1002/sim.6113.
[34] P. Czodrowski, "Count on kappa," *J Comput Aided Mol Des*, vol. 28, no. 11, pp. 1049–1055, 2014, doi: 10.1007/s10822-014-9759-6.
[35] S. Y. Ng, K. M. Lim, C. P. Lee, and J. Y. Lim, "Sentiment analysis using DistilBERT," in *2023 IEEE 11th Conference on Systems, Process & Control (ICSPC)*, IEEE, Dec. 2023, pp. 84–89. doi: 10.1109/ICSPC59664.2023.10420272.
[36] Y. Yang, C. Miller, P. Jiang, and A. Moghtaderi, "A case study of multi-class classification with diversified precision recall requirements for query disambiguation," in *SIGIR 2020 - Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, Association for Computing Machinery, Inc, Jul. 2020, pp. 1633–1636. doi: 10.1145/3397271.3401315.