Environment Sentiment Analysis of Bali Coffee Shop Visitors Using Bidirectional Encoder Representations from Transformers (BERT) and Generative Pre-trained Transformer 2 (GPT2) Model

Ni Putu Widya Yuniari[1,*], Ni Made Satvika Iswari[2,], I Made Surya Kumara[3]

[1,3]*Indonesia Faculty of Engineering and Planning, Warmadewa University, Denpasar 80239, Indonesia*

[2]*Faculty of Information Technology and Design, Primakara University, Denpasar 80226, Indonesia*

(Received: July 22, 2024; Revised: August 18, 2024; Accepted: September 13, 2024; Available online: October 15, 2024)

Abstract

Bali is one of the provinces with the most abundant natural and cultural wealth in Indonesia. One commodity that supports it is coffee. Bali Coffee is not only a gastronomic identity, but also a cultural identity which makes it have added value to be developed into various business lines. One business derivative that is quite promising is a coffee shop. However, these favorable conditions also need to be maintained to ensure good quality reaches consumers. One thing that can do is analyze reviews from customers. One of the most popular methods is Sentiment Analysis. This technique allows business to analyze customer reviews on social media. It can be a feedback to maintaining and improving quality and good relationships with customers. This research aims to create a machine learning model to analyze customer reviews at several coffee shops in Bali which are divided into three labels, namely: positive, negative and neutral. The methods used are: scraping, cleaning, stopword removal, embedding, undersampling, and modeling. The algorithms used are Bidirectional Encoder Representation from Transformer (BERT) and Generative Pre-trained Transformers (GPT). The performance metrics used in this research are precision, recall, accuracy and loss. This research succeeded in creating a sentiment analysis model for coffee shop customers in Bali. The BERT model obtained an accuracy value of 78% without undersampling with a loss in the 10th iteration of 0.27. Meanwhile, the BERT model with undersampling obtained an accuracy value of 32.85% with a loss in the 10th iteration of 0.16. The GPT2 model without undersampling gets an accuracy of 78% with a loss in the 10th iteration of 0.25. Meanwhile, the GPT model with undersampling obtained an accuracy value of 32.85% with a loss in the 10th iteration of 0.15.

Keywords: BERT, GPT2, NLP, Sentiment Analysis

1. Introduction

The tourism industry was expanding rapidly in Bali. One of the provinces that has the most unique culture and tourism background in Indonesia especially after Covid19 pandemic [1]. Considering that this sector contributed 5.5% to GDP and contributed 16,910 million USD to the country's foreign exchange in 2019. These advantages make Bali a province that is very rich in natural and cultural resources in the tourism sector. The tourism and creative economy sectors are the main contributors to the economy of Bali. According to the Bali Provincial Bureau of Statistics (BPS), one of the sub-sectors that most contribute is accommodation, food, and beverages, especially coffee shop [2].

Coffee is one of the leading commodities and culinary industries in Bali. The types of coffee produced are Arabica coffee and Robusta coffee. Most Balinese coffee is produced from coffee plantations in the Kintamani district. This coffee was then called Kintamani coffee. [3]. Coffee is a very valuable part of Balinese culture, both traditional culture and modern culture which refers to hustle culture. As a traditional culture, coffee is often served as part of rituals offered to ancestral spirits e.g. at the Tumpek Uduh and Tumpek Wariga rituals. However, as a modern culture, hustle culture has taken Balinese coffee even further. The habit of drinking coffee in an artisan style in a café with an aesthetic ambiance has become a part of the culture of modern Balinese society. This habit was especially brought by Generation Y and Generation Z [4]. Hustle culture brings the habit deep into society to the point that where it can form a collective habit. For one example, it can be seen from the popularity of the term work from coffee shop (wfc) which has popularized the nomad style of working from coffee shop to coffee shop which is very popular with young people

***Corresponding author: Ni Putu Widya Yuniari (putu.widyayuniari@warmadewa.ac.id)**
DOI: https://doi.org/10.47738/jads.v5i4.302

ISSN 2723-6471

today. This popularity ultimately continues to grow and makes the coffee shop a business that promising for investment for the future.

According to research by the Global Agricultural Information Network, it shows that the projected domestic consumption in 2019/2020 will reach 294,000 tonnes, an increase of around 13.9% compared to consumption in 2018/2019 which reached 258,000 tonnes. In 2021, domestic consumption of Indonesian coffee is estimated to increase again to 370,000 tons [5]. This escalation is far above the growth of global coffee consumption, which is only 8% by the same term [6]. According to research conducted independently by Toffin and in collaboration with Mix magazine, it shows that the number of coffee shops in Indonesia in August 2019 reached more than 2,950 outlets, an increase of almost three times compared to 2016 which was only around 1,000 [7]. This escalation has forced coffee shop owners in Bali to use various methods to attract customers, one of which is by providing a comfortable and aesthetic place. Ultimately setting a new standard for an artisan coffee shop.

Nowadays, Balinese people and tourists visiting Bali have various choices and types of coffee shops that they can visit in Bali. Of course, each coffee shop has its advantages and disadvantages. As a business entity, we understand that existing strengths must be perfectly maintained, while deficiencies must be improved as much as possible. One method that can be used to find out the advantages and disadvantages of a business entity is to analyse reviews given by customers. This analysis is called sentiment analysis [8]. Currently, customers can provide direct reviews of a business entity through various channels such as Trip Advisor, Google Maps, Google Local Business, and various other channels. Customers can also create an independent review in the form of a blog, thread, tweet, or personal review.

This analysis helps business owners understand customer needs, criticism, and suggestions to continue to grow and develop their business. This technique can easily be conducted manually, but imagine if a business entity had so many reviews that it was impossible to analyse them one by one. In this case, a machine learning model is needed that can understand the meaning of human opinion and create a model to predict and determine whether a review is positive or negative.

One of the latest studies on sentiment analysis was conducted by Magdaleno et al. This study used a Large Language Model (LLMs) with a transfer learning approach to analyse customer reviews. The model used is GPT (Generative Pre-trained Transformer). This study obtained an MAE value of 0.27, MSE of 0.26, R-squared of 0.69 and MAPE of 0.10% [9]. Another study on sentiment analysis of coffee shop customer reviews was conducted by Mawadati et al. The method used was Data Mining. This research aims to increase customer satisfaction. The technique used was to extract customer reviews into several labels that are not supervised first. This means that the machine was left to work to find out the cluster of sentiment that is given by the customer [4].

Chandradev et al conducted an additional investigation with the objective of developing a sentiment review analysis model for examining hotel customer reviews in Bali post-pandemic. Sentiment review analysis created by model to analyse hotel customer reviews in Bali after the pandemic. The algorithm used in this research is smallBERT. The approach used was supervised learning where the data was labelled first before the training process. This research obtained an accuracy of 91.40%, precision of 90.51%, recall of 90.51%, and F1 Score of 90.51% [8].

Branco et al conducted another research in the field of sentiment analysis. This study tried to analyse the reviews of several Portuguese restaurants that were available on the Zomato application. This study used a supervised approach where the dataset was labelled first before entering the training process. This research uses the BERT algorithm. In this research, the base BERT model was modified with an ensemble technique thereby increasing the model accuracy from 80% to 84% [10].

Kheiri and Karimi had utilize the Large Language Model (LLM) with a transfer learning approach to analyse sentiment data. The algorithm used was GPT which has been improved to become GPT-Turbo. To compare, this study also uses other GPT variants, including Ada, Babbage, Curie, Davinci, Embedding + Xgboost, and Embedding + RF. This study succeeded in obtaining accuracy on the GPT-Turbo model of 97.3%, recall of 91.98%, and F1 Score of 94.26% [11]. Another study using a transfer learning approach was conducted by Sayeed et al. This study uses BERT as a tokenization algorithm. The approach used was unsupervised learning, where BERT will learn the data structure carry out tokenization, and then separate it based on the tokens produced. This study concludes that BERT has significant

 ISSN 2723-6471

performance for handling complex data compared to other models. The BERT model succeeded in providing good performance, namely a precision score of 86%, recall of 78%, and F1 Score of 82% [12].

Another study conducted used a deep learning approach. However, this study is less popular at the moment because there are already various models that have better performance and efficiency. One of them was conducted by Xiaoyan et al who used GloVe as a feature extraction algorithm and CNN and LSTM as classification algorithms [13]. Another study was conducted by Amalia and Winarko who used Word Embedding as a feature extraction algorithm and CNN as a classification algorithm [14]. Another study was conducted by Cahyaningtyas et al who used ABSA as a feature extraction algorithm and compared various deep learning algorithms such as RNN, LSTM, GRU, BiLSTM, and CNN [15]. Another study was conducted by Hossai by combining CNN and LSTM [16].

Magdaleno et al [9] research contributed to the selection of GPT2 as the algorithm to be used. This selection is also supported by Kheiri and Karimi's [11] research which obtained significant results using the GPT model. Research by Chandradev [8], Branco [10] and Sayeed [12] provides recommendations for a comparison algorithm, namely BERT. The big idea about sentiment analysis techniques and methodology refers to Mawadati's research [4]. Meanwhile, several other studies provide insight into algorithms that enrich research but are not used due to novelty issues.

This research will create a sentiment analysis model by comparing the performance of BERT and GPT2 algorithms on customer reviews obtained from Google Maps of several coffee shops in Bali. The novelty of this research lies in the data scraping method and data sources used. If various previous studies used Twitter data crawling, this research tried to use Google Maps Review data scraping with different characteristics. In each model, data balancing would be performed using random-undersampling. This technique is also new in this research and can be compared with previous studies. This technique was chosen based on the results of a literature review, where the oversampling technique was feared to produce artificial tokens that were never owned by the data. This study then compares model performance between models with and without data balancing. This research aims to create and find out the model that has the best performance in determining the sentiment of a customer's review of a coffee shop. This modeling can simplify and speed up the analysis process so that it can help coffee shop business owners adapt and make decisions to advance their business.

This research needs to be conducted to provide a better alternative for sentiment analysis by using more varied sentiment channels than previous studies. So that the resulting analysis is more comprehensive (not limited to the Twitter channel) and can complement previous research.

2. Literature Review

2.1. Natural Language Processing (NLP)

Data processing technique that uses machine learning to extract, process, and interpret data in the form of text in a native language [17]. In simple terms, NLP is defined as the ability of a computer to understand human language using certain computational algorithms [8]. During the last decade, the application of machine learning techniques has led to achievement of higher accuracy in many types of NLP applications [18]. NLP tries to uncover the structure and meaning of text and convert it into numerical data. Examples of using NLP include Google Assistant, Apple Siri, Open AI Chat GPT, and Google Bard (which has been upgraded to Gemini). NLP enables these platforms to understand user intent and communicate naturally like real people.

NLP works by changing every word into numeric tokens that are both meaningful and easy to understand by machines. This process is called tokenization [19]. From these tokens, the machine will realize a particular function, pattern, or distribution of the texts. From that function, the machine then gets a conclusion [20]. The tokenization process has several techniques including TF-IDF, NER, Random Tokenizer, etc. The newest method that is most often used in various studies recently is transfer learning, which is reusing models that have been previously trained (pre-trained) and then retrained to improve performance on different data [10].

Conventional tokenization techniques in the Python programming language usually use the NLTK (Natural Language Toolkit) library. This library not only provides tokenization functions, it also provides stopword removal, lexical analysis, and lemmatizer functions [20]. In advanced tokenization processes, transformer libraries are usually used.

 ISSN 2723-6471

This library stores a lot of pre-training models that are ready to use. Examples include BERT & GPT. These two models have changed the NLP paradigm from text-based training to Large Language Models (LLMs) which already store various tokens with a much larger size [21]. NLP can help users quickly capture meaning, keywords, and relevant data within a text message. The application of NLP helps readers to quickly identify stress words [22].

2.2. Sentiment Analysis

Sentiment analysis is a method used to extract, interpret, and reason for a user's sentiment and analyze it to find out the contents of a person's thoughts, feelings, and also a person's assessment or review of a case or issue on the free opinion column area [8]. In machine learning, sentiment analysis is also recently named opinion mining [13]. Sentiment analysis is one manifestation of NLP (Natural Language Processing). This method concentrates on extracting and interpreting subjective data from various sources. The main goal is to understand the sentiment or stance of a speaker or writer on a particular topic or to see contextual polarities in an opinion [11].

Sentiment analysis is usually used to perform clustering or predict whether a sentiment has a positive, negative, or neutral sentiment [13]. But sentiment analysis can work further than that. Sentiment analysis can be used to cluster a thread or an opinion into fact/hoax classes. This technique can also be used to track the first user who shares this opinion. Sentiment analysis can also work using an unsupervised approach where a computational model will form a function based on word entities and group them into several clusters. Another thing that sentiment analysis can do is extract each text to find out the words most frequently used by users or the terminology that is popular in a particular case or issue. The evolution of sentiment analysis over time is characterized by three important phases: the Lexicon-based era, the Machine Learning (ML) era, and the Transformer model era based on Large Language Models (LLMs) [11]. The method used to conduct sentiment analysis, especially in the case of data classification, is as follows:

2.2.1. Data Collection

Data collection is a process for searching and collecting data. The most popular techniques to use are data crawling and scraping. The crawling process can be conducted using software, API, or pseudocode. However, in many cases, users need to have an authorization token to be able to crawl the data. For example, when we crawl the Twitter data. Another approach that can be taken is scraping. Scraping is the process of extracting hardcoded data. That is, by creating a code that forces the program to search for information in a certain area and retrieve it. An example is when taking YouTube comment data or reviews from Google Maps [23].

2.2.2. Data Preparation

Data preparation is the process of preparing data before entering the processing stage. Some of the most popular techniques used in this process include case-folding and data cleaning. Case folding is the process of changing the entire text to lowercase, while cleaning is done to clean data from certain words or characters that are not expected, for example: symbol characters, emojis, non-alpha-numeric characters, website URLs, hashtags, mentions, etc. [20].

2.2.3. Stopword Removal

Stopword Removal is the process of removing unimportant words from sentences because they are very common and do not have substantial meaning. Examples are: a, an, the, is, are, etc. These words need to be removed, the aim is to avoid bias in the sentence structure and create a reliable model. Stopword Removal is the process of removing unimportant words from sentences because they are very common and do not have substantial meaning. For example: a, an, the, is, are, etc. These words need to be removed, to avoid bias in the sentence structure and create a reliable model [20].

2.2.4. Stemming (Lemmatizer)

Stemming (lemmatizer) is a process for changing non-standard words into their standard form. For example, changing 'went' or 'gone' to 'go'. The aim is to reduce bias and increase data homogeneity. This is because 'go', 'went', and 'gone' actually have the same meaning but will have different tokens. So, if this is allowed to happen there will be two patterns that are not identical. Therefore, all non-standard words need to be normalized into their standard form [20].

2.2.5. Tokenization

Tokenization is the process of separating sentences per word and making them into tokens where each word is indexed with a certain number. This index functions to prevent repetition, as well as to count the number of words. There are several methods for tokenizing, including a tokenizer from NLTK, a tokenizer from Tensorflow, a tokenizer from Wordcloud, and the tokenizer function from Transformers.

2.2.6. Embedding

Machines can understand a sentence by converting the sentence into a numeric vector first and then discovering patterns. Word embedding is the process of converting words in the form of alphanumeric characters into vector form. The vector formed can be an array in Python or a Tensor for PyTorch. There are several methods for doing word embedding in Python, including Word2Vec, TF-IDF, and using the Embedding function from BERT.

2.2.7. Padding

Padding is the process of equalizing the size or dimensions of the vector that is produced in the word embedding process. In this process, a vector that has more dimensions than the specified dimensional threshold will be truncated at that limit. On the other hand, vectors that have dimensions less than the specified dimensional threshold will be completed using the number zero (0). This process aims to equalize the dimensions of the vector with the dimensions of the layers used so that it can go through an artificial neural network process. Various methods can be used to conduct padding in Python, including doing it manually by fiddling the vectors or using the padding function from TensorFlow. On the other hand, if the program was conducted using Transformers, padding is usually conducted automatically after the Embedding process. So, there is no need to create additional functions.

2.2.8. Modeling

Modeling is the process of creating a sentiment analysis model. This process involves machine learning algorithms to create a function that can understand patterns in a text vector and translate them into certain insights. This model will then go through a training and evaluation process to provide the ability to understand new sentences. Models commonly created in sentiment analysis include positive-negative prediction, hoax detection, hate speech detection, question and answer, chatbot, text generation, essay marking, etc. The algorithm used will be adapted to various needs.

Sentiment analysis techniques are widely applied for various purposes, including the tourism industry. This technique enables developers to extract and analyze data quickly and precisely to create a user interaction model and perform many other analyses. The impact can be seen in the development of a user-centric industry that is able to improve user experience in several sectors. Data from sentiment analysis can be used both as insight and recommendations to create and implement new policies and improve user-centric promotional strategies.

2.3. Large Language Model (LLMs)

This model that has been trained and enriched with a giant corpus [11]. Models such as BERT, T5, and GPT have transformed NLP models or functions into LLMs. These models have a much larger number of corpora than models such as GloVe and TF-IDF. This corpus has been trained and can still be improved using transfer learning. LLMs models have also been enriched with various languages and various needs. For example, to enrich BERT with criminal corpora, CrimeBERT is usually used. To enrich BERT with health corpora, BioBERT is usually used. This technology then became the most advanced technology which created various new technologies such as Bard, ChatGPT, and Google Gemini. The terms most frequently used in LLMs are transfer learning and fine-tuning. Transfer learning is the process of retraining a previously trained model to improve model performance on new data. Then the trained model is used for certain purposes. Meanwhile, fine-tuning is the process of re-tuning the model architecture and retraining to get more optimal results [10].

2.4.Imbalanced Learning

Imbalance learning is a method for balancing data on categorical data. This type of data is separated into certain categories or classes that indicate data clusters. In some cases, the amount of data in each class or category is not balanced. Sometimes there is one class that has a very large amount of data while another class has a very small amount of data. If the gap is too high, it is feared that it will cause training bias which will result in a decrease in the performance

 ISSN 2723-6471

of the machine learning model. An indication of this can be seen in the confusion matrix results which tend to predict data only in one particular class [24].

Imbalance learning can be conducted using two approaches, namely: oversampling and undersampling. Oversampling is an approach in imbalance learning that conducts sample generation in classes that have less data. Generation is conducted randomly by finding the mid value between two pieces of data in the same class and determining this value as new data. This technique ensures that the new data remains in the same distribution, so it does not change the data distribution in that class. One of the most popular methods is SMOTE Oversampling. Sample generations continue to be multiplied until all classes have the same amount of data [25].

In contrast to oversampling, undersampling is an approach in imbalance learning that aims to reduce the amount of data to a certain level. In this case, until all classes have the same amount of data. The data selected for deletion is usually random and does not have any particular standards. This approach is often considered better than the oversampling approach because this approach does not generate new data that is not real data. This method is also preferred on the LLMs models provided by Transformer. This is because these models have 3 different pieces of data which makes it more complicated to generate new data. Token data, which is usually in text form, cannot simply be regenerated, so it is against the rules to change it to numeric first. In this case, an undersampling approach is preferred. This approach may reduce the amount of data and change its distribution, but that will not be a problem as long as we have enough data with high homogeneity [26]. Imbalance learning in Python is usually conducted using the imbalanced-learn or imb-learn libraries. However, it can also be done manually or by creating your automatic function. Using the imbalance learning library does look easier, but it is not dynamic for large data using LLMs. In these cases, the approach of creating your own more dynamic functions is preferable [24].

2.5. Undersampling

Undersampling is an approach in imbalance learning that equalizes the amount of data in each class by reducing the data in other classes that have a larger amount of data. There are several methods commonly used in undersampling, including Random Undersampling, Cluster Undersampling Technique (CUST), and Trainable Undersampling.

Random Undersampling is an undersampling method that deletes data randomly without using a particular function or approach. The data that is deleted is data in the class that has more data up to a predetermined amount, usually the same as the amount of data in the class that has the least data. This technique is the technique most widely used in undersampling, this is because it is simple to use without having to use other functions or formulas first. This technique does have drawbacks, the data distribution can change during the data deletion process. However, this can be avoided by having sufficient data with high homogeneity. This technique is also the most preferred in LLMs because of its simplicity. This technique can be done manually or by creating your automatic function using Python at any step and on any form of data. So, it is not related to certain rules [27].

Cluster Undersampling Technique (CUST) is an undersampling technique that creates clusters first in each data class, and then simultaneously removes data in each cluster. This technique is more complicated than random undersampling, this is because we have to create a clustering algorithm first before deleting the data. But in some cases, this technique works better, this is because the clustering process carried out will maintain the data distribution at certain centroid points so that it does not change the data distribution too much. This technique is usually used if we only have a little data or have data with low homogeneity [28].

The last technique commonly used is Trainable Undersampling. This technique uses a machine learning model to perform undersampling. Machine learning models are used to choose which data should be deleted in certain classes. Usually, the data that is deleted is data that has a low-class probability, so the hope is that apart from deleting the data, it can also improve performance. But so far, this method is the most difficult to use. This is because we have to create a model without using undersampling first and with that model the data is deleted and re-trained with new data [29].

2.6. BERT

BERT is a deep learning model designed to process natural language processing. An NLP architecture that uses attention mechanisms to replace recurrent networks and can capture relationships between contextually distant words. BERT utilizes encoder-decoder techniques to find out the relationship between words in a sentence and produces a

 ISSN 2723-6471

representation for each word in the sentence as output so that it can improve model performance on complex sequential tasks in NLP [8].

BERT was originally published by Google researchers as: "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding" [30]. BERT builds on recent work in pre-training contextual representations, including Semi-supervised Sequence Learning, Generative Pre-Training, ELMo, and ULMFit. However, unlike previous models, BERT is a strictly bidirectional first language representation. The architectural visualization can be seen in figure 1.

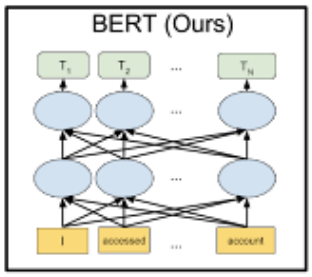

Figure 1. BERT Architectural Representation [31]

A visualization of the BERT neural network architecture compared to previous state-of-the-art contextual pre-training methods is shown below. Arrows show the flow of information from one layer to the next. The green box at the top shows the final contextual representation of each input word.

BERT is a representation of the encoder function of the Transformer model. BERT is the first fine-tuning-based representation model to achieve state-of-the-art performance on a large set of sentence-level and token-level tasks and outperforms many task-specific architectures [31]. BERT is currently provided open source by Transformers. To use the BERT model, we can import the model from Transformers. Currently, BERT has provided various models for purposes such as Sentence Classification, Question Answering, BERT for Next Sentence Prediction, and BERT for Multiple Choice & Token Classification. Bert is also available for various fields of science such as CrimeBERT for criminology and BioBERT for health. Apart from that, BERT is also available in various languages such as IndoBERT etc.

2.7. GPT2

GPT is a family of neural network models that use a transformer architecture. GPT was published by Open AI and academically published by Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever as "Language Models are Unsupervised Multitask Learners" [32]. GPT is an important advancement in machine learning that supports generative applications such as ChatGPT. This model gives users the ability to create text and content that resembles human creation. The concept is that GPT converts text into vectors, then studies and understands it, then creates a function from these insights and uses it to generate other content.

The GPT architecture may look simpler than the BERT model, where the direction of information flow is carried out in the next text. This approach is called the one-way model. An illustration can be seen in figure 2.

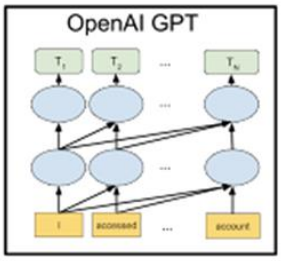

Figure 2. GPT Architectural Representation [31]

One-way models are trained efficiently by predicting each word conditioned on previous words in the sentence. This will allow the indirectly predicted word to “see itself” in a multi-layer model. To use the GPT2 model we can import

 ISSN 2723-6471

it using Transformers. Currently, GPT2 has many functions for various purposes such as Sequence Classification, Question Answering, and Token Classification. The functions that GPT2 has in the context of NLP do appear to be fewer than BERT, this is because GPT is not specifically for processing NLP, but uses NLP to generate other content. However, its performance can still be relied on for carrying out NLP operations such as sentiment analysis.

3. Methodology

This research is generally divided into four stages, namely: data collection, data labeling, data analysis, and evaluation. The methodology of this research can be seen in figure 3. Initially customer data review of local businesses, such as coffee shops was gathered through Google Maps website. These reviews typically include detailed feedback and ratings (on a scale of 1 to 5 stars) provided by customers who have visited the coffee shops. For a project focusing on Bali coffee shops, reviews from Google Maps can offer insights into customer satisfaction, service quality, product offerings, and the overall atmosphere of the shops.

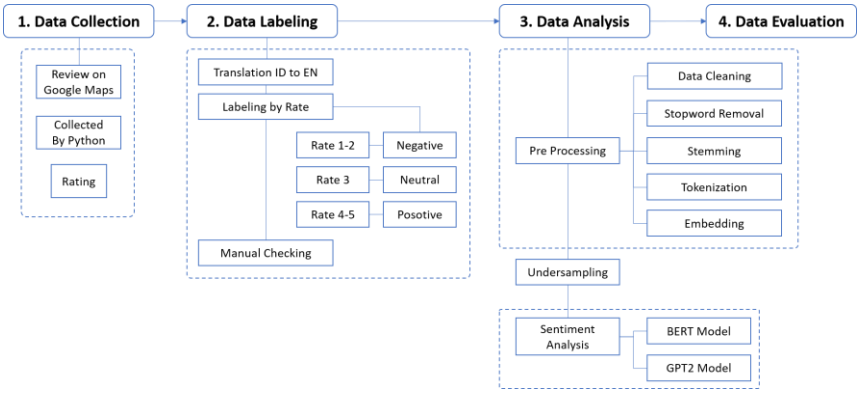

Figure 3. Research Methodology

The second phase of this research is data labeling, a vital part of sentiment analysis. This step involves assigning sentiment labels (like positive, negative, or neutral) to a set of textual data. Typically, human annotators are required to read and evaluate each text, determining its sentiment according to established guidelines or criteria. The data labeling process in sentiment analysis begins by giving human annotators a representative sample of pre-labeled data.

3.1. Data Collection

The data collection process was conducted by scraping reviews of several coffee shops in Bali, Indonesia through Google Maps Review. Data collection was conducted using Scrapping Python on the Google Maps Review Page from several predetermined coffee shops. The list of cafes used can be seen in table 1.

Table 1. Café List

No	Café Name
1	Blend Cafe
2	Gangga Coffee
3	Otokafe
4	Home Cafe
5	Kafe
6	Mudra Cafe
7	Tropical View
8	Little Talks

The list of cafes in table 1 was chosen because the popularity of these cafes lies not only in interior design, but also in the cultural richness brought through gastronomic entities. In this phase, data also goes through an automatic deletion process for repetitive data. From the cafes in table 1, 1336 data were obtained consisting of 1051 (78,67%) positive

 ISSN 2723-6471

review data, 194 (14,52%) negative review data, and 91 (6,81%) neutral review data. The comparison percentage can be seen in figure 4. One of the scraped reviews can be seen in table 2.

Table 2. Example Review

Review	Label
We ate very, very well. The number is relatively large. The fruit juices are amazing and the smoothie bowls are simply amazing! An address to remember.	Positive
Thai on the menu with bus prawns. I only found one shrim :)))	Neutral
Even though it looked very pretty, the taste was so bland that it was strange. don't know how something can look so good but taste like that. not cold at all & the fruit is warm. never come back. Instagram story worthy	Negative

The percentage diagram of data distribution based on labels or sentiment categories can be seen in figure 4.

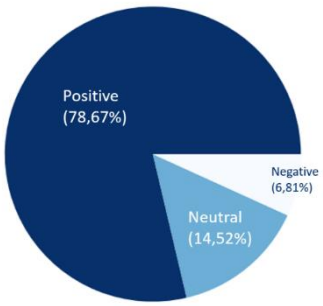

Figure 4. Review Class Percentage

From the diagram in figure 4 it can be concluded that 78.76% of the data are positive reviews, 6.81% are negative reviews and the remaining 14.52% are neutral reviews.

3.2. Data Preparation

Data preparation in this research was divided into 3 stages, namely: label encoding, missing value analysis, and data cleaning. The purpose of this process was to prepare data before entering the further process. Label encoding was the process of changing labels from text to numeric. It must be conducted because machines cannot read text directly, but must be converted into numerical. The labels that are encoded include: positive becomes '1', neutral becomes '2', and negative becomes '3'. Missing value analysis was a process to find out data that was missing or non-existent (NaN) or had a Null value during the data cleaning process. This analysis has been begun by searching for missing data. Then in the second analysis, these missing data will be removed. This needs to be conducted before entering the next process because the process of embedding the BERT model and GPT will experience bias and training failure if there is still missing data. However, if in the first analysis, no missing data is found, then the research can continue in the next process. During the missing data test, testing is carried out on 'review' and 'label' data. The result was that no missing data was found in both data. This means that cleaning does not need to be done and research can proceed to the next stage. The final stage in the data preparation process was cleaning the data from unrelated characters, including non-alpha-numeric characters, website (URL), tagging, mentions, and encoding characters. In this process, all review data would have been standardized into lowercase. The results of the data-cleaning process can be seen in table 3.

Table 3. Cleaning Result

Step	Text
Base	Very nice place, the smoothie bowls look amazing and taste amazing too. Recommend this place
Cleaning	very nice place the smoothie bowls look amazing and taste amazing too recommend this place

From table 3 it can be seen that all characters are now lowercase and there are no non-alphanumeric characters.

 ISSN 2723-6471

3.2.1. Stopword Removal

The stopword removal process aims to remove words that are too common and unimportant. The library has been used to conduct this process was the NLTK. Stopwords, such as "the," "and," and "a," often do not contribute significantly to the meaning of a text. By removing them, we can reduce the dimensionality of the data and improve the performance of machine learning models. NLTK provides a pre-built stopword list for English, but it can also be customized to include or exclude specific words based on the requirements of the task. For example, in a domain-specific application, industry-specific terms might be excluded from the stopword list. The results of the stopword removal process can be seen in table 4.

Table 4. Stopword Removal Result

Step	Text
Cleaning	very nice place the smoothie bowls look amazing and taste amazing too recommend this place
Stopword Removal	nice place smoothie bowls look amazing taste amazing recommend place

3.2.2. Stemming (Lemmatizer)

Stemming (Lemmatizer) was a process to return words to their original form. The library has been used to carry out this process is the NLTK. This process is crucial for text analysis tasks as it helps to reduce the dimensionality of the data by grouping together words with similar meanings. For example, the words "run," "running," and "ran" can all be stemmed to "run," which can improve the accuracy of text classification and clustering algorithms. The results of the steaming process can be seen in table 5.

Table 5. Stemming Result

Step	Text
Stopword Removal	nice place smoothie bowls look amazing taste amazing recommend place
Stemming	nice place smoothie bowl look amazing taste amazing recommend place

3.2.3. Tokenization

The tokenizing process in this research used the BERT Tokenizer function and GPT2 Tokenizer function. The corpus used is 'bert-base-uncased' for BERT and ‘microsoft/DialogRPT-updown’ for GPT2. The 'bert-base-uncased' corpus consists of 30,522 words divided into 768 token classes and the ‘microsoft/DialogRPT-updown’ consists of 50,257 words and 1,024 token classes. It can be seen that GPT2 has richer tokens, so the hypothesis leads to GPT2 having better performance than BERT. Tokenization in natural language processing that involves breaking down text into smaller units, such as words or subwords. BERT and GPT2 use different tokenization strategies, which can impact the quality of the resulting tokens. In this research, the hypothesis is that the richer tokenization of GPT2 will lead to better performance in downstream tasks, such as text classification and question answering

3.2.4. Embedding

The embedding process in this research used the encoding function from BERT and GPT2. This process converts word tokens into vectors in Tensor form. Then the padding process is carried out so that the vector length becomes 32 for the entire data. This process produces 3 pieces of data that will be used during the training process, including Input IDS & Encode IDS, Attention Mask and Labels. BERT and GPT2 use different embedding techniques, which can affect the quality of the learned representations. Segment IDs are used to indicate the boundaries between different sentences or segments in the input text, which is particularly important for tasks like question answering or text summarization.

3.3. Undersampling

Undersampling is a technique used to address class imbalance problems in machine learning, where one or more classes have significantly fewer examples than others. The undersampling process used in this research is random undersampling by creating an automatic function using Python. Random undersampling works by randomly reducing data on data that has high labels. This reduction continues until all labels have the same amount of data. From the

 ISSN 2723-6471

results of figure 4 above, it can be concluded that after the undersampling process, each class has the same amount of data. The results can be seen in figure 5.

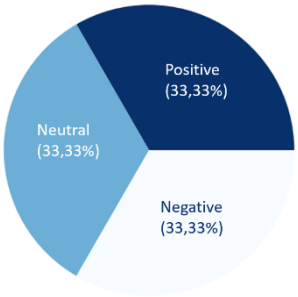

Figure 5. Review Class Percentage After Undersampling

3.3.1. BERT Model

The BERT model used in this research is the BERT Model Uncase from Transformers. The language used is English. This model is then fine-tuned using a new dataset to produce a better model.

3.3.2. GPT2 Model

The GPT2 model used in this research is the 'microsoft/DialogRPT-updown' model. The difference is, that the padding process is carried out only up to 20 data, resulting in a leaner Tensor. The language used is English. This model is then fine-tuned to produce a model that has better performance.

3.3.3. Training

Before entering the training process, the data will be divided into training data and testing data first with a division of 50% for training data and 50% for testing data. The training process is carried out using the AdamW optimizer for both BERT and GPT2. The learning rate used is 1 x 10-4. The batch size used is 32, meaning that the training process for each iteration will be divided into 32 batches. The training process was carried out in 10 iterations (epochs).

3.4. Evaluation Metric

The training performance measurement metric used was a loss. The loss represents errors generated during the training process. The loss value calculation is carried out automatically by the model so that in this research the loss value is generated directly during the training process. The loss function used is cross entropy as shown in formula (1) below:

$$\text{Loss} = -\sum_{c=1}^{N} y_c \log(p_c) \quad (1)$$

In the BERT and GPT2 model, the predicted token of the masked word is passed to a softmax layer which converts the masked word's vector into another embedding. Then we can calculate the cross-entropy loss between the input vector and the one we got after the softmax layer.

The next measurement metric was the confusion matrix. The confusion matrix is a measurement matrix that compares predicted data with actual data in each class. An example of a confusion matrix can be seen in figure 6.

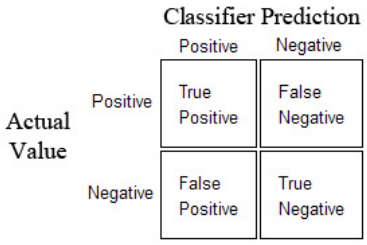

Figure 6. Confusion Matrix Example [33]

 ISSN 2723-6471

The confusion matrix gives us 4 pieces of data, namely True Positive (positive data that was successfully predicted), True Negative (negative data that was successfully predicted), False Negative (positive data that failed to predict), and False Positive (negative data that failed to predict).

The next metric used in this research was accuracy. Accuracy states how accurately the model predicts data whether positive, negative, or neutral. To calculate the accuracy value of the confusion matrix, Formula (2) can be used.

$$\text{cc} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \tag{2}$$

The next metric used in this research was precision. Precision compares the number of positive data that are successfully predicted divided by the number of data that are classified as positive. To calculate the precision of the confusion matrix, Formula (3) can be used.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{3}$$

The last metric used in this research was recall. Recall is a comparison between the amount of positive data that was successfully predicted divided by the amount of data that was positive. Recall states how well the model predicts positive data. To calculate recall from the confusion matrix, Formula (4) can be used.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{4}$$

4. Results and Discussion

4.1. BERT Without Undersampling

In BERT testing without using undersampling, the training and validation loss results were obtained at 10 epochs as in figure 7. From the training results, it was found that the training process succeeded in reducing the loss value in the training data from 0.3569 to 0.0195. Meanwhile, in the validation data, there was also a very significant decrease, from 0.3639 to 0.2456. The confusion matrix produced in this section can be seen in figure 8. This experiment resulted in an accuracy of 78%, precision of 78% and recall of 78%. The loss calculation in this model produces a loss of 0.27.

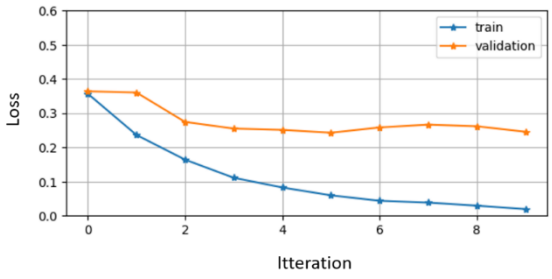

Figure 7. BERT Without Undersampling Training Result

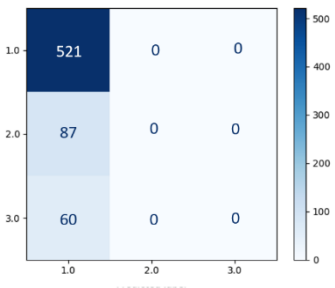

Figure 8. Confusion Matrix BERT Without Undersampling

4.2. BERT With Undersampling

In BERT tested using undersampling, the training and validation loss results were obtained at 10 epochs as in figure 9. From the training results, it was found that the training process succeeded in reducing the loss value in the training data from 0.1701 to 0.0266. Meanwhile, in the validation data, there was also a very significant decrease, from 0.1794 to 0.1182. The confusion matrix produced in this section can be seen in figure 10. This experiment resulted in an accuracy of 32.85%, precision of 32.85% and recall of 32.85%. The loss calculation in this model produces a loss of 0.16.

ISSN 2723-6471

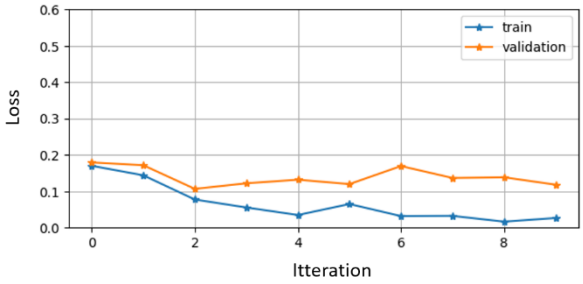

Figure 9. BERT With Undersampling Training Result

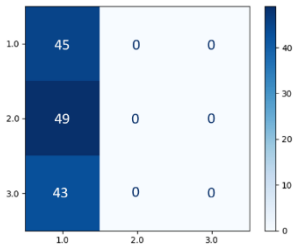

Figure 10. Confusion Matrix BERT with Undersampling

4.3. GPT2 Without Undersampling

In GPT2 tested without using undersampling, the training and validation loss results were obtained at 10 epochs as in figure 11. From the training results, it was found that the training process succeeded in reducing the loss value in the training data from 0.9399 to 0.0139. Meanwhile, in the validation data there was also a very significant decrease, from 0.3834 to 0.2588. The confusion matrix produced in this section can be seen in figure 12. This experiment resulted in an accuracy of 78%, precision of 78% and recall of 78%. The loss calculation in this model produces a loss of 0.25.

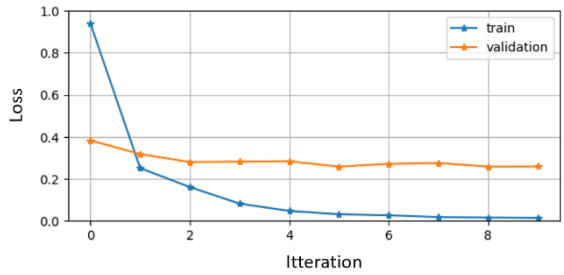

Figure 11. GPT2 Without Undersampling Training Result

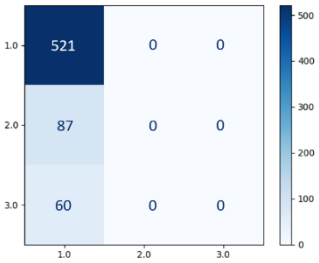

Figure 12. Confusion Matrix GPT Without Undersampling

4.4. GPT2 With Undersampling

In GPT2 testing with using undersampling, the training and validation loss results were obtained at 10 epochs as in figure 13. From the training results, it was found that the training process succeeded in reducing the loss value in the training data from 0.3822 to 0.0128. Meanwhile, in the validation data there was also a very significant decrease, from 0.3926 to 0.1929. The confusion matrix produced in this section can be seen in figure 14. This experiment resulted in an accuracy of 32.85%, precision of 32.85%, and recall of 32.85%. The loss calculation in this model produces a loss of 0.15.

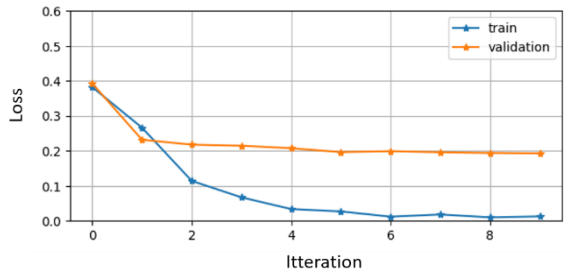

Figure 13. GPT2 With Undersampling Training Result

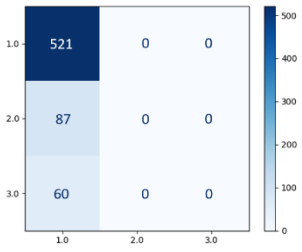

Figure 14. Confusion Matrix GPT2 With Undersampling

 ISSN 2723-6471

4.5. Discussion

From the results of the discussion above, a performance matrix can be created from several test scenarios. The summary can be seen in table 6 below.

Table 6. Comparison Matrix

Model	Balancing Data	Accuracy	Precision	Recall	Loss
BERT	Without Undersampling	78%	78%	78%	0.27
BERT	With Undersampling	32.85%	32.85%	32.85%	0.16
GPT2	Without Undersampling	78%	78%	78%	0.25
GPT2	With Undersampling	32.85%	32.85%	32.85%	0.15

From table 6 above, it can be seen that the BERT and GPT models have the same accuracy, precision, and recall performance both on data with and without data balancing. In the model with balancing data, the accuracy value produced by BERT is 78%. This value is the same as the value produced by GPT2. Meanwhile, the accuracy value for the model without data balancing produced by BERT is 32.85%. This value is also the same as the value produced by GPT2. This may be caused by the amount of data being too little. Moreover, after dividing it into training data and testing data, it is also reduced using random undersampling. Too little data indicates that there is too little learning material to make the model perform well. Too little data also indicates that the model may experience the phenomena of too high homogeneity or too high heterogeneity. These two phenomena both cause bias. Homogeneity that is too high can be seen from the prediction results which are only collected on one label, while heterogeneity that is too high can be seen from the loss value which is also high.

Based on Table 6, the resulting loss values tend to be low. The value gets better with each iteration. This means that the data can't experience extreme heterogeneity. However, based on the confusion matrix, it can be seen that the prediction data is collected on only one label. This indicates that the data we have is too homogeneous. This also explains why the performance of accuracy, precision, and recall after undersampling decreases drastically. That is because the model gets the same prediction results in each class. Because the accuracy, precision, and recall values are the same for the BERT and GPT models, it is necessary to analyze other performance metrics to compare model quality, namely Loss. Based on Table 6, the Loss value for the GPT2 model is smaller than the BERT model. This means that based on this metric, the GPT2 model performs better than the BERT model. From this loss analysis, it was also concluded that the undersampling process, although it reduces the accuracy value, also reduces the loss value.

5. Conclusions

Based on the experimental results, it can be concluded that the BERT and GPT 2 models work very well. Even though the precision, recall, and accuracy values tend to be the same, from the loss values it can be concluded that the GPT 2 model provides better performance. From the results of this experiment, it can also be concluded that the random undersampling process is able to reduce the loss value but also reduces the accuracy value. Sentiment analysis plays a crucial role in the development of tourism, particularly within its vibrant coffee industry in Bali, which embodies rich gastronomic and cultural identities. By analyzing customer reviews using this technique, businesses can gain insights into customer perceptions and responses, thereby enhancing service quality. This study underscores the importance of employing advanced machine learning models like BERT and GPT2 to analyze customer feedback from diverse coffee shops across Bali. The research findings demonstrate that sentiment analysis effectively categorizes customer reviews into positive, negative, and neutral sentiments, with both models achieving high accuracy. Techniques such as undersampling are explored to optimize model performance despite potential reductions in accuracy. Consequently, sentiment analysis serves as a valuable tool for Bali's coffee industry to elevate customer experiences and develop effective marketing strategies, thereby contributing to sustainable and competitive tourism development.

6. Declarations

6.1. Author Contributions

Conceptualization: N.P.W.Y., N.M.S.I., and I.M.S.K.; Methodology: N.M.S.I.; Software: N.P.W.Y.; Validation: N.P.W.Y., N.M.S.I., and I.M.S.K.; Formal Analysis: N.P.W.Y., N.M.S.I., and I.M.S.K.; Investigation: N.P.W.Y.; Resources: N.M.S.I.; Data Curation: N.M.S.I.; Writing Original Draft Preparation: N.P.W.Y., N.M.S.I., and I.M.S.K.; Writing Review and Editing: N.M.S.I., N.P.W.Y., and I.M.S.K.; Visualization: N.P.W.Y.; All authors have read and agreed to the published version of the manuscript.

6.2. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

6.3. Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

6.4. Institutional Review Board Statement

Not applicable.

6.5. Informed Consent Statement

Not applicable.

6.6. Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] A. R. A. Putri, P. G. Fadhila, and A. Furqan, "Tourism Impact on Economic Growth in Bali," *Services for Science and Education*, vol. 11, no. 6, pp. 94–101, Jun. 2023.

[2] BPS, "Indikator Pertumbuhan Ekonomi Bali Triwulan IV-2023," Badan Pusat Statistik Provinsi Bali, Bali, 2023.

[3] N. L. Suastuti, "Tourist Satisfaction Towards Arabica Coffee at Catur Village Kintamani Bangli - Bali," *Advances in Economics, Business and Management Research*, vol. 111, no. 2018, pp. 85–89, 2018.

[4] Mawadati, W. Ustyannie, A. H. Wibowo and R. A. Simanjuntak, "Analysis of Yogyakarta Coffee Shop Visitor Reviews to Increase Customer Satisfaction Using Sentiment Analysis," in *International Conference on Engineering Management and Sustainable Innovative Technology*, vol. 2014, no.1, pp. 30-39, 2024.

[5] A. Fitch et al., "The Coffee Compromise: Is Agricultural Expansion into Tree Plantations a Sustainable Option?," *Sustainability*, vol. 14, no. 5, pp. 3019-3033 Mar. 2022.

[6] International Coffee Organization, "Annual Review Coffee Year 2021/2022," International Coffee Council, London, United Kingdom of Great Britain and Northern Ireland, 2022..

[7] A. O. P and Dr. N. Samatan, "Coffee shop Marketing Communication Strategy 'Neng Sumi' in retainning comsumers during PSBB in Jakarta," *International Journal of Communication, Management and Humanities*, vol. 1, no. 2020, Nov. 2020.

[8] P. P. Rokade and A. K. D, "Business Intelligence Analytics using Sentiment Analysis-A survey," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 9, no. 1, pp. 613–620, Feb. 2019.

[9] D. Magdaleno, M. Montes, B. Estrada and A. Ochoa-Zezzatti, "A GPT-based Approach for Sentiment Analysis and Bakery Rating Prediction," *Advances in Computational Intelligence. MICAI 2023 International Workshops (MICAI 2023)*, vol. 2024, no. Jan., pp. 61-76, 2024.

[10] Branco, D. Parada, M. Silva, F. Mendonça and S. S. Mostafa, "Sentiment Analysis in Portuguese Restaurant Reviews: Application of Transformer Models in Edge Computing," *electronics*, vol. 13, no. 589, pp. 1-20, 2024.

[11] K. Kheiri and H. Karimi, "SentimentGPT: Exploiting GPT for Advanced Sentiment Analysis and its Departure from Current Machine Learning," Utah State University, Utah, 2023.

[12] M. S. Sayeed, V. Mohan and K. S. Muthu, "BERT: A Review of Applications in Sentiment Analysis," *HighTech and Innovation Journal*, vol. 4, no. 2, pp. 453-462, 2023.

[13] L. Xiaoyan, R. C. Raga and S. Xuemei, "GloVe-CNN-BiLSTM Model for Sentiment Analysis on Text Reviews," *Hindawi Journal of Sensors*, vol. 2022, no. 1, pp. 1-12, 2022.

[14] P. R. Amalia and E. Winark, "Aspect-Based Sentiment Analysis on Indonesian Restaurant Review Using a Combination of Convolutional Neural Network and Contextualized Word Embedding," *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, vol. 15, no. 3, pp. 285 – 294, 2021.

[15] S. Cahyaningtyas, D. H. Fudholi and A. F. Hidayatullah, "Deep learning for aspect-based sentiment analysis on Indonesian hotels reviews," *Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control Journal*, vol. 6, no. 3, pp. 239-248, 2021.

[16] N. Hossain, M. R. Bhuiyan and Z. N. Tumpa, "Sentiment Analysis of Restaurant Reviews using Combined CNN-LSTM," *in 11th ICCCNT 2020*, vol. 11, no. 1, pp. 1-5, 2020.

[17] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu and P. Kuksa, "Natural Language Processing (almost) from Scratch," *Journal of Machine Learning Research*, vol. 1, no. 1, pp. 1-48, 2000.

[18] H. Florentina Hristea , C. Cornelia, Preface to the Special Issue "Natural Language Processing (NLP) and Machine Learning (ML)—Theory and Applications", *Mathematics*, vol. 2022, no. 10, pp. 2481-2489, 2022.

[19] A. Thawani, J. Pujara, F. Ilievski, and P. Szekely, "Representing Numbers in NLP: a survey and a vision," *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, vol. 2021, no. 1, pp. 644–656, Jun. 2021.

[20] D. Yogish, T. N. Manjunath, R. S. Hegadi, "Review on Natural Language Processing Trends and Techniques Using NLTK" *Recent Trends in Image Processing and Pattern Recognitionvol*, vol. 1037, no. 1, pp. 589-606. 2019.

[21] H. W. Kim and H. Chang, "From Words to Numbers: Getting Started with Text Analysis for Applied Social Scientists," *Business Communication Research and Practice*, vol. 3, no. 2, pp. 122-129, 2020.

[22] P. M. Mah, I. Skalna, and J. Muzam, "Natural Language Processing and Artificial Intelligence for Enterprise Management in the Era of Industry 4.0," *Applied Sciences*, vol. 12, no. 18, p. 9207, Sep. 2022.

[23] Yunelfi, R. Putri, "DarkWeb Crawling using Focused and Classified Algorithm." *[CEPAT] Journal of Computer Engineering: Progress, Application and Technology, [S.l.]*, vol. 1, no. 02, pp. 1-6, aug. 2022. ISSN 2963-6728.

[24] M. A. Khder, "Web Scraping or Web Crawling: State of Art, Techniques, Approaches and Application," *Int. J. Advance Soft Compu. Appl*, vol. 13, no. 3, pp. 144-168, 2021.

[25] S. Santhiappan and B. Ravindran, "Class Imbalance Learning," *Reconfigurable and Intelligent Systems Group, Department of Computer Science and Engineering, IIT Madras*, vol. 2017, no. 1, 2017. doi: 10.34048/2017.1.F1.

[26] J. H. Joloudari, A. Marefat, M. A. Nematollahi, S. S. Oyelere and S. Hussain, "Effective Class-Imbalance learning based on SMOTE and Convolutional Neural Networks," *Appl. Sci. ,* vol. 13, no. 6, pp. 1-12, 2022.

[27] X.Y. Liu, J. Wu and Z.-H. Zhou, "Exploratory Undersampling for Class-Imbalance Learning," *IEEE Transactions on System, Man and Cybernetics*, vol.39, no. 2, pp 539–550, 2008.

[28] M. A. N. M. Yusuf and M. C. Untoro, "Evaluate of Random Undersampling Method and Majority Weighted Minority Oversampling Technique in Resolve Imbalanced Dataset," *IT Journal Research and Development (ITJRD)*, vol. 8, no. 1, pp. 1-13, 2023.

[29] R. A. Sowah, M. A. Agebure, G. A. Mills, K. M. Koumadi and S. Y. Fiawoo, "New Cluster Undersampling Technique for Class Imbalance Learning," *International Journal of Machine Learning and Computing*, vol. 6, no. 3, pp. 205-214, 2016.

[30] M. Peng, Q. Zhang, X. Xing, T. Gui, X. Huang, Y.-G. Jiang, K. Ding and Z. Chen, "Trainable Undersampling for Class-Imbalance Learning," *The Thirty-Third AAAI Conference on Artificial Intelligence (AAAI-19)*, vol. 33, no. 1, pp. 4707-4714, 2019.

[31] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, vol. 1, no. 2019, pp. 4171–4186, Jun. 2019, doi: 10.18653/v1/N19-1423..

ISSN 2723-6471

[32] G. Yenduri, "GPT (Generative Pre-Trained Transformer)— A Comprehensive Review on Enabling Technologies, Potential Applications, Emerging Challenges, and Future Directions," *in IEEE Access*, vol. 12, no. 1, pp. 54608-54649, 2024.

[33] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," *OpenAI Blog*, vol. 1, no. 8, pp. 9-17, 2019..

[34] J. Banda, R. Angryk, and P. Martens, "Steps Toward a Large-Scale Solar Image Data Analysis to Differentiate Solar Phenomena," *Solar Physics*, vol. 288, no. 1, May 2013, doi: 10.1007/s11207-013-0304-x.