CHAPTER I
INTRODUCTION

1.1 Background

Bali, as one of the most popular tourist destinations in the world has become an icon of Indonesian tourism known for its natural beauty, cultural richness, and the hospitality of its people (Suniadewi, 2024). Bali's popularity has led to rapid growth in tourism. The rapid growth of tourism can be seen from various indicators, one of which is the increase in the number of tourist visits (Aliansyah & Hermawan, 2021). The increasing trend in the number of visits can be observed in Figure 1.1 which shows data on tourist visits to Bali from 2019 to 2023.

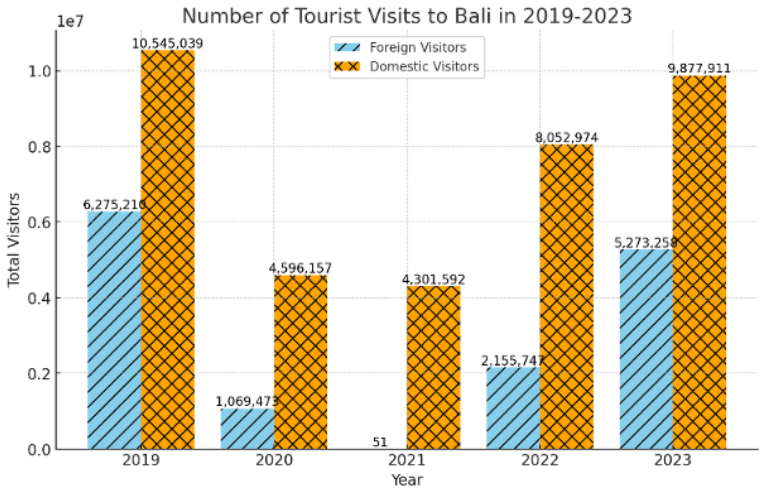

Figure 1.1 Number of Tourist Visits to Bali in 2019-2023

Source: Badan Pusat Statistik Provinsi Bali (2025)

According to data from the Badan Pusat Statistik Provinsi Bali (2025), accessed on February 4, 2025, tourist arrivals to Bali plummeted in 2020 due to the COVID-19 pandemic, reaching a low of just 51 foreign visitors in 2021. With the easing of travel restrictions, tourism began to recover in 2022 and continued to rise through 2023, nearing pre-pandemic levels and reaffirming Bali's position as a world-leading tourist destination.

Rapid tourism generally brings positive impacts, such as boosting foreign exchange and strengthening service industries that support economic growth (Yamamoto et al., 2021). As a top global destination, Bali significantly contributes to Indonesia's economy through local revenue and employment opportunities (Adhi et al., 2025). However, rapid tourism growth also poses sustainability challenges, particularly in resource and infrastructure management. Southern Bali, for instance, is facing gentrification, and overtourism or overcapacity issues with traffic congestion and overcrowded attractions like Uluwatu Temple, which hosts around 6,000 tourists daily despite limited seating (NusaBali, 2024). Reports from *Fodors* and *CNA* (2024) highlight how overtourism has led to uncontrolled development and reduced comfort, making Bali less appealing and more congested than before.

The rapid growth of tourism in Bali has sparked diverse public opinions, especially on TikTok, which is the world's most popular social media that allows users to create, watch, and share short videos, as well as discuss through the commenting feature (D'Souza, 2025). In research conducted by (Nurrozan & Zulfebriges, 2025) on the Relationship between Social Media Exposure to TikTok and Interest in Traveling, proves that exposure to TikTok social media, especially through text, image, audio, and video elements, plays a major role in increasing interest in traveling. Indonesia became the country with the most TikTok users in the world as of July 2024, reaching 157.6 million users (Fatika, 2024). While TikTok effectively promotes Bali's tourism through viral content, it also exposes negative impacts such as environmental damage, traffic congestion, and cultural pressure. Comment sections often reflect heated debates, revealing public concerns. However, the absence of a sentiment monitoring system can delay government responses, potentially worsening social conflict and damaging Bali's image as a sustainable and culturally rich tourism destination. Therefore, leveraging sentiment analysis technology to monitor public opinion on TikTok is crucial for timely and informed policymaking. The large volume of data on TikTok prompted the need for a web scraping-based approach to access information systematically. In this study, TikTok data will be collected using Apify, a platform for web scraping and automation, allowing users to extract, process, and utilize web data efficiently (Jason, 2024).

To provide a more comprehensive understanding of this issue, this research will conduct a sentiment analysis of the opinions on TikTok. The purpose of this sentiment analysis is to identify patterns of positive, negative, and neutral sentiments contained in various content and comments related to rapid tourism in Bali. Sentiment analysis can be done with various classification methods, one of which is the Bidirectional Encoder Representations from Transformers (BERT) method, especially IndoBERT which is a pre-trained deep learning model designed to handle data in Indonesian (Koto et al., 2020). In research conducted by (Riyadi et al., 2024) and (Fransiscus & Girsang, 2022) it was found that the use of BERT-based methods, such as IndoBERT, showed better performance compared to other methods such as Support Vector Machine and Naïve Bayes. These findings show that IndoBERT is more effective in analyzing sentiment in Indonesian language compared to these traditional methods.

In sentiment analysis, there is a data labelling process before the data is used. The data labelling process is a very crucial stage that is usually carried out before the data is used, because it greatly affects the accuracy of the classification results. Usually, data labelling is done by experts in the field or annotators, but this process often faces considerable challenges, especially in terms of availability and involvement of competent annotators. As a solution, one approach that is widely used is lexicon-based automatic labelling, such as the InSet Lexicon which is an Indonesian-specific lexical dictionary that has been proven to provide higher results in classifying sentiment on Indonesian-language datasets compared to other lexicon-based such as VADER (Fathoni et al., 2024). Labelling with Lexicon-based offers high efficiency and consistency as the process is rule-based and can be done on a large scale. However, despite the advantages, labelling with InSet has limitations such as the context of the sentence is not always well captured difficulty in handling ambiguity. Meanwhile, manual labelling done by humans allows a deeper understanding of the context but requires more time and resources especially if the dataset is large and is prone to subjectivity because sentiment interpretation can differ between labelers even on the same data.

Therefore, this research aims to conduct a comparison between automatic labelling approaches using InSet Lexicon and manual labelling by experts, both of

which are integrated into the IndoBERT model. This integration is designed to evaluate the effect of labelling quality on model performance in sentiment classification. By harnessing the power of transformer-based models that have proven to excel in understanding linguistic context, this research is expected to identify the most effective and efficient data labelling method in the context of Indonesian sentiment analysis, specifically related to public perception of tourism development in Bali. In addition to contributing to the improvement of classification accuracy, the results of this comparison are also expected to make a practical contribution in determining the appropriate data labelling strategy in the development of sentiment analysis systems. Specifically, this research also aims to measure the effectiveness of the InSet approach in the data labelling process, both in terms of the accuracy of the resulting model and the efficiency of time and resources in the annotation process.

1.2 Problem Identification

Based on the background previously described, several main issues can be identified as follows:

1. Rapid tourism growth in Bali has led to a number of detrimental impacts, such as overtourism, environmental degradation, exploitation of natural and cultural resources, and a decline in the quality of tourist experiences due to excessive crowding. Although rapid tourism growth also brings positive impacts, such as economic growth through mass tourism, provides high returns on the increase of foreign exchange and service industries that play a role in improving the economy, contributes greatly to the local revenue, and is also a major source of employment for the local population
2. Public sentiment toward Bali's rapid tourism growth has become increasingly visible through social media, especially TikTok. The platform is widely used to express public concerns related to environmental degradation, overcrowding, and cultural shifts caused by the influx of tourists. However, the lack of real-time sentiment analysis has resulted in delays in policy response and risked worsening public dissatisfaction and Bali's tourism image. Therefore, there is a need to analyze public sentiment using a model that can handle both large data volume and contextual accuracy.

3. From a methodological standpoint, one of the key problems in this research lies in identifying an effective and efficient labelling method for sentiment classification. Although IndoBERT has been proven to deliver strong performance in analyzing Indonesian-language text, it still depends on the availability of high-quality labelled data. Manual labelling by human annotators is known for its accuracy in capturing linguistic subtleties and context, such as sarcasm, cultural references, or emotional tone, but it is time-consuming, resource-heavy, and difficult to scale for large datasets. On the other hand, automatic labelling using the InSet Lexicon offers a fast and scalable alternative, yet it often lacks contextual understanding and may mislabel nuanced expressions. These limitations present a critical challenge how to ensure labelling quality that supports accurate sentiment classification at scale. Therefore, the research seeks to address this issue by evaluating both manual and automatic labelling approaches when integrated with IndoBERT, to determine the most effective method for understanding public sentiment regarding Bali's rapid tourism growth.

1.3 Problem Statement

Based on the identification of these issues, the research questions for this study are formulated as follows:

1. How is the performance comparison of IndoBERT integrated with automatic labelling using InSet Lexicon and manual labelling by human annotators in sentiment classification of Bali's rapid tourism growth?
2. How are the sentiment patterns regarding Bali's rapid tourism growth revealed through sentiment analysis using IndoBERT with the best-performing sentiment labelling method?

1.4 Research Objective

Referring to the problem formulation that has been described, the following are the objectives of this research:

1. To analyze the performance comparison between IndoBERT integrated with automatic labelling using InSet Lexicon and manual labelling by human annotators in sentiment classification of Bali's rapid tourism growth.

2. To analyze sentiment patterns related to Bali's rapid tourism growth using IndoBERT with the most effective sentiment labelling approach.

1.5 Scope of Research

In order to make the research to be carried out more focused, the following is the scope of research in this study:

1. This research focuses on sentiment analysis of public opinion about Bali's rapid tourism growth based on data taken from the TikTok social media platform.
2. The dataset used consists of TikTok video comments primarily written in Indonesian. Comments written in a mixture of Indonesian and English are included, with the English parts translated into Indonesian. However, comments written entirely in foreign languages, such as English, are excluded from the analysis.
3. The data used includes a collection of comments with keywords and hashtags such as "Bali", "Pariwisata Bali", "Pertumbuhan Pariwisata Bali" as well as hashtags #bali, #pariwisataBali, and #pertumbuhanPariwisataBali.
4. Comments from TikTok videos posted within the last two years were collected, specifically covering the period from January 2023 to December 2024.
5. This research includes the development of a simple web that presents sentiment analysis results on public opinion regarding tourism growth in Bali. The web aims to visualize sentiment classification outcomes obtained from the integration of the IndoBERT model with the labelling method that demonstrated the best performance. Therefore, this study does not discuss web development principles in depth, nor does it include system testing activities such as usability testing, performance evaluation, or user interface assessment, as these are beyond the scope of this research.

1.6 Research Significance

Referring to the problem formulation and objectives of this research that have been described previously, the following are the benefits that can be obtained through this research, as follows:

1. Theoretical Significance

a. This research contributes to the development of Natural Language Processing (NLP) studies, particularly in sentiment analysis using Indonesian language. By comparing the performance of IndoBERT, InSet Lexicon, and manual label integration, this research expands the understanding of the effectiveness of the methods in capturing sentiment contextually.
b. This research provides empirical validation of IndoBERT performance in the complex, informal and culturally rich context of Indonesian language, such as in tourism narratives in Bali. It adds theoretical insights into the ability of transformer-based models to handle Indonesian texts.
c. By comparing lexicon-based, machine learning, and manual labelling approaches, this research contributes to the theory of combining or integrating approaches in improving sentiment classification accuracy.

2. Practical Significance
    a. The results of this sentiment analysis can be utilized by stakeholders such as local government, the Tourism Office, and industry players to understand public and tourist perceptions of the impact of Bali's tourism growth, both in terms of economic, social, and environmental.
    b. This research can be a real case study in the utilization of AI technology for data-based decision making, as well as a reference for students and academics in similar research.