1 Introduction

Every day, many users review anything on the internet including a product and other users can witness it because anything written on the internet will become public consumption. That said, other people's opinions are necessary to be a source of information for someone's belief in judging a product to be purchased online [1]. One site that provides user reviews of a product about food is Zomato. Zomato — formerly Foodiebay — is a directory information service for multinational restaurants or food and delivery services [2]. Zomato was created in 2008 and launched in 2010 as a platform for finding restaurants, reading and writing user-generated reviews, ordering food delivery, and ordering tables to make payments at dine-in at restaurants [3] and also provides images of dishes from available restaurants and reviews from customers who have visited [2]. Until this moment, Zomato is one of the top startups in India and operates in around 23 countries including Indonesia, valued at USD 1 Billion (2016) [4] and has 32.1 million monthly active users worldwide in FY21 [5] and also ranked first in the IDN Times version of Indonesia's best food review application [6].

Even though there have been many reviews about a restaurant product on Zomato, many reviews are not convincing or biased, whether good or bad. These unreliable reviews can be disastrous for some. Seeing from the perspective of potential buyers, it won't be apparent to convince potential buyers to buy the product or not. Meanwhile, from the perspective side of the restaurant owner, it can become a boomerang for the restaurant so that it can't promote its restaurant. Recent empirical studies have shown that the volume and influence of online consumer reviews significantly affect product sales [7]. So this case raises the need for an automated model for product rating [8]. The automated model used in this research is sentiment analysis. As one of the applications of natural language processing (NLP), Sentiment analysis extracts sentiment information from text input given through classes, for example, positive or negative [9]. To process the food review into several parts, meaning to find out whether the review belongs to positive or negative sentiment, is to use document-level sentiment analysis. At the document level, the sentiment predicate is taken from the entire document, with input as a sentence and output as a class [10]. Sentiment analysis is used to understand the emotional tone behind the person writing [11].

This study aims to build a sentiment analysis model of food reviews on the Zomato website. The process stages start from preprocessing, feature extraction using Term Frequency-Inverse Document Frequency (TF-IDF), and classification using the K-Nearest Neighbor (KNN) algorithm. The KNN method is used because KNN is a classification method with the best accuracy of 96.8% when paired with the TF-IDF feature extraction [12]. TF-IDF also produced best accuracy of 80.0% combined with KNN when compared with other feature extractions [16]. TF-IDF will be used to find out how often a word appears because the classification of text data using the K-Nearest Neighbor (KNN) classification method requires the layout distance of each existing data. The dataset used was obtained from the Zomato website with 1953 reviews data on two restaurants in Bandung, Indonesia, namely One Eighty Coffee and Se'i Sapi Lamalera restaurants. The language used in this dataset is Indonesian. There are two labels in the dataset, namely positive 1004 data and negative 949 data. Each labeled review is limited to focusing on reviews in terms of food, not including atmosphere, cleanliness, or service. The dataset will be splitted into two parts, namely train data and test data with ratio of 60:40, 70:30, and 80:20. The points that need to be underlined in this research are as follows:

1. The Zomato balanced dataset used in this study are dataset that had never been studied before.
2. For data normalization preprocessing stage, we made a word normalization dictionary containing 233 words that need to be normalized.
3. For meaningless word removal preprocessing stage, we made a meaningless word dictionary containing 282 words that need to be removed.
4. We evaluated the effect of preprocessing stage, split data ratio comparison, and comparison between euclidean distance and manhattan distance calculation method.

The following section that will be discussed in the research is section 2 regarding the study of literature related to research. Furthermore, in section 3, namely an explanation of the proposed model, the results and analysis discussed in chapter 4 and the last, chapter 5, conclusions from the research that has been evaluated.