Compare VGG19, ResNet50, Inception-V3 for review food rating

Andrew1)*, Handri Santoso2)
1)2)3) Magister Teknologi Informasi, Universitas Pradita
1)email@student.pradita.ac.id, 2)handri.santoso@pradita.ac.id

Submitted : Apr 16, 2022 | Accepted : Apr 17, 2022 | Published : Apr 18, 2022

Abstract: The food industry is undergoing a phase of very good improvement, where business actors are experiencing very rapid growth. Creative ideas are many and creative on several social media. When an online business is growing rapidly, many managers in the food sector market their products through online media. So it is quite easy for customers to place orders via mobile. Especially during the COVID-19 pandemic, where a ban on gatherings has become a government recommendation for many food business actors to sell online. Since then, almost all food industry players have made their sales online. There are many advantages of doing business online. The food served is in the form of pictures that attract market visitors so that it can create its own charm. Food is just a click away to order, and the order comes. No need to queue and everything has been delivered to the ordered goods. After the ordered goods arrive, the customer reviews the food or drink. Because customer reviews are the result of customer ratings. The result of the review is one of the sentiment analyses, which in this study is in the form of a review of the images available on the display marketplace. The method used is Convolutional Neural Network. The dataset will be extracted features and classifications. The research will do a comparison using VGG19, ResNet50, and Inception_v3. Where the accuracy of VGG19 = 96.86; Resnet50 : 97.29; Inception_v3 : 97.57.

Keywords: Sentiment Analysis; Food Rating; Image Review; Convolutional Neural Network, Dataset.

INTRODUCTION
Ordering food is now very easy, just use your finger, food comes quickly without waiting in line. This method is the implementation of the digital marketing concept which has become a culture in the country. Food providers only concentrate on the product to be produced, namely the best food. There are now many marketplaces that provide stalls for them to sell. With a business model like this it is easy to sell products such as food.

Sentiment analysis is one of the popular branches of Natural Language Processing with the aim of expressing feelings into certain texts (Lopamudra Dey et al., 2016). Sentiment analysis has also been explored more deeply by many researchers in various languages, including Indonesian. Research on sentiment analysis depends on a particular domain or specific science, sentiment political science (Elghazaly et al., 2016), science in healthcare (Alayba et al., 2017), cybersecurity science (AlRowaily et al., 2015), customers provide reviews about service companies (Almuqren & Cristea, 2016), and reviews related to economics (Yang et al., 2020).

The purpose of sentiment analysis is to classify public opinion into positive class, neutral class, and negative class in all fields of science, including education, marketing, politics, and economics. Attention because people are more communicative on social media, especially Twitter (Shelke & Korde, 2020). In fact twitter contains a variety of large tweets or short posts that come from various users and encourage people to communicate and there is an exchange of information and exchange of ideas. Therefore, most researchers from Indonesia use Twitter data as a dataset to extract public opinion in obtaining important information (Prastyo et al., 2021).

Various datasets related to food datasets, and there have been many studies that discuss problems in food. Starting from foods related to raw materials for meat, fish, various foods used by China, Japan, Western and from Asian countries. The dishes are served in an image-changing dataset. There are many millions of images and more than thousands of classes for various foods that are presented in the image dataset.

*name of corresponding author

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.
Table 1. Dataset for food from various datasets

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Year</th>
<th>Classes/Images</th>
<th>Public</th>
</tr>
</thead>
<tbody>
<tr>
<td>Food-975 (Zhou &amp; Lin, 2016)</td>
<td>2016</td>
<td>975/37,785</td>
<td>x</td>
</tr>
<tr>
<td>Food500 (Merler et al., 2016)</td>
<td>2016</td>
<td>508/148,408</td>
<td>x</td>
</tr>
<tr>
<td>Food11 (Singla et al., 2016)</td>
<td>2016</td>
<td>11/16,643</td>
<td>x</td>
</tr>
<tr>
<td>UNICT-FD1200 (Farinella et al., 2016)</td>
<td>2016</td>
<td>1,200/4,754</td>
<td>√</td>
</tr>
<tr>
<td>Food524DB (Ciocca et al., 2018)</td>
<td>2017</td>
<td>524/247,636</td>
<td>√</td>
</tr>
<tr>
<td>Vegfru (Hou et al., 2017)</td>
<td>2017</td>
<td>292/160,000</td>
<td>√</td>
</tr>
<tr>
<td>FoodX-251 (Teh et al., 2018)</td>
<td>2019</td>
<td>251/158,846</td>
<td>√</td>
</tr>
<tr>
<td>ISIA Food-200 (Min et al., 2019)</td>
<td>2019</td>
<td>200/197,323</td>
<td>√</td>
</tr>
<tr>
<td>ISIA Food-500 (Min et al., 2020)</td>
<td>2020</td>
<td>500/399,726</td>
<td>√</td>
</tr>
<tr>
<td>Food2K (Min et al., 2021)</td>
<td>2021</td>
<td>2,000/1,036,564</td>
<td>√</td>
</tr>
</tbody>
</table>

Some of the datasets used by research related to food are as shown in table 1. Some of the existing datasets are private and some of them are public, the datasets are shared for further research. The purpose of this study is a research for sentiment analysis using a food image dataset. The image datasets will later vary, there are positive, negative and neutral datasets, which cause feelings of pleasure, sadness, normal, happy, angry, surprised and so on.

LITERATURE REVIEW

This literature review will describe the studies that discuss sentiment analysis. Previous studies are the basis of research. The following in table 2 is a list of previous studies.

Table 2. Previous research on the topic of sentiment analysis

<table>
<thead>
<tr>
<th>Author</th>
<th>Topic</th>
<th>Advanced</th>
<th>Disadvanced</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Aaputra, 2019)</td>
<td>E-Wallet Sentiment Analysis on Google Play Using Naive Bayes Algorithm Based on Particle Swarm Optimization</td>
<td>Sentiment Analysis using Rapidminer tools with Naive Bayes Classifier and Particle Swarm Optimization feature selection.</td>
<td>The model has not been said to be good, if it has not been tested with dataset testing, and has not been developed with the addition of images from the actors of sentiment analysis.</td>
</tr>
<tr>
<td>(Laurensz &amp; Eko Sediyono, 2021)</td>
<td>Analysis of Public Sentiment on Vaccination in Efforts to Overcome the Covid-19 Pandemic</td>
<td>Discussion on public sentiment analysis for the covid-19 vaccine. Using datasets from social media Twitter, Support Vector Machine algorithm, and Naive Bayes.</td>
<td>The model has not been said to be good, if it has not been tested with dataset testing, and has not been developed with the addition of images from the actors of sentiment analysis.</td>
</tr>
<tr>
<td>(Lia Farokhah, 2020)</td>
<td>Convolutional Neural Network Implementation for Classification of Emotion Intensity Variations in Dynamic Image Sequence</td>
<td>Discussion about sentiment analysis using Deep Learning with Convolutional Neural Network algorithm, as the main focus on Facial emotion recognition.</td>
<td>Regarding sentiment analysis, using the Convolutional Neural Network only focuses on facial emotion recognition, it doesn't show which object is the result of facial changes.</td>
</tr>
<tr>
<td>(Juwiantho et al., 2019)</td>
<td>Image Sentiment Analysis Using Deep Convolutional Neural Network With Concept Features</td>
<td>Discussion about sentiment analysis using image dataset and text dataset from Twitter social media, using</td>
<td>The results obtained from image-based sentiment analysis do not explain the sentiment analysis of</td>
</tr>
</tbody>
</table>
Convolutional Neural Network algorithm faces and comments or object sentiment analysis more clearly so this research can be used as a reference.

(Susanto, 2021) Sentiment Analysis and Topic Modeling in Online Learning in Indonesia Through Twitter The discussion about sentiment analysis with the topic of online learning using the Twitter social media dataset using Fine-grained sentiment analysis The method used in this sentiment analysis is not explained in detail the classification using machine learning or deep learning, but only explains the calculations manually.

From previous studies such as in table 2, research on sentiment analysis mostly uses sentences with several machine learning algorithms such as nave Bayes, support vector machines and others. The state of the art in this study is to use an image dataset using a deep learning algorithm, namely convolutional neural networks and combined with the sentiment analysis method to produce positive, negative and neutral food output. This research is a complement to previous studies, where this study uses an image dataset. This research is not looking for shortcomings of previous research.

METHOD

The research method proposed for this research can be seen in fig. 1, where all images will be downloaded from several public image datasets and private images. The private image comes from the image taken by the camera. The public image is from Table 1. Dataset for food from various datasets. Then the images are grouped into the train and test datasets. The training dataset is grouped into a 5-star rating folder, 4-star rating folder, 3-star rating folder, 2-star rating folder, and 1-star rating folder. Likewise with the grouping of the testing dataset.

The training dataset and testing dataset have been managed with their respective folders, then to the Convolutional Neural Network stage. The Convolutional Neural Network stage is the process of performing feature extraction and classification. There are 5 classes in this classification. 1 star class for very dissatisfied reviews, 2 stars for dissatisfied reviews, 3 stars for moderate reviews, 4 stars for satisfied reviews, and 5 stars for very satisfied reviews.

*name of corresponding author

Fig. 1 Proposed Image Sentiment Analysis methodology.

Source : researcher property

This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.
Convolutional Neural Network

Convolutional Neural Networks (CNN) algorithm is a very common algorithm in Deep Learning. This algorithm is generally used to classify objects. For example, to classify 2 classes, namely true class, and false class, to classify 5 classes, namely very satisfied review, satisfied review, moderate review, dissatisfied review, and very dissatisfied review.

Inside CNN there are several types such as VGG19, Resnet50, Inception_V3, MobileNet, and others. Functions in CNN are feature extraction and classification. The architecture of VGG19 can be seen in Figure 2.

![Convolutional Neural Network](source)

Image as input is done by convolution, for example with matrix Conv, max polling is done until the input image becomes a bottleneck. Then it is used as a vector and continued as input from the neural network. Neural network, all neurons are connected to the output neurons with fully connected, it will produce a classification. The process of convolutional Conv + ReLu then polling, Conv + ReLu then polling, Conv + ReLu then polling becomes a bottleneck, this process is called a feature map. The Bottleneck is rendered into a Flatten Layer.

Furthermore, by performing feature extraction into classification and prediction. Classification in this study there are 5 classes rating 1, rating 2, rating3, rating 4, and rating 5.

**VGG19**

In 2012 AlexNet made improvements to the Convolutional, neural network. Actually, VGG is an upgrade from AlexNet and created by the Visual Geometry Group at the Oxford campus and is better known as VGG. There is a VGG algorithm with 16 layers which contains 13 Convolutions, 2 fully connected, and softmax layer output. In addition, there is VGG19 which contains 16 convolution layers and 3 full connected layer. A comparison of VGG16 and VGG19 can be seen in fig. 3.

![VGG16 vs VGG19](source)

*This is an Creative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.*

*name of corresponding author*
**Resnet-50**

ResNet-50 is the architecture of the Convolutional Neural Network that utilizes a deep learning process using a residual network.

![ResNet-50 Architecture](image)

Figure 4. shows the detailed architecture of the ResNet50-vd network (ABDULFATTAH et al., 2021), (Qiu et al., 2021). The entire network in figure 2 on the left consists of a parent module, four residual modules, and a fully connected neural network layer. The numbers 32, 64, 256 in figure 2 represent the number of convolution channels. The parent module consists of three 3×3 convolutions and max pooling, and uses stride = 2 in the first convolution and Max pooling to achieve downsampling. Therefore, the output feature map of the master module is half the size of the input, and the number of channels is 64. The modules from stage 1 to stage 4 contain one block1 and several blocks2. It can be seen from Figure 2b that block1 consists of two paths to form a downsampling block. On the left side is a bottleneck structure consisting of three convolutions, which are used to learn new features. On the right side is a structure consisting of convolution and AvgPooling, which is used to process the input into the same size and scale as the output of the bottleneck structure. However for stage2 to Stage4, block1 with parameter stride = 2 to scale the feature map size in half. Block2, like block 1, consists of two paths. The difference is that the structure on the right side of block 2 is a shortcut connection that forms the residual module with the lower structure. The output feature map sizes from trunk to Stage4 are [7, 14, 28, 56, 112]. The number of convolution channels from stem to Stage 4 is [64, 256, 512, 1024, 2048]. Therefore, the final output feature map of Stage4 is 7×7×2048. The last FC layer of the model uses the mean pooling layer and the fully connected layer.

**Inception_V3**

Inception V3 includes a deep learning model using Convolutional Neural Networks, for image classification. Inception V3 is an improvement over Inception V1 which was discovered by GoogLeNet around 2014.
process of Inception includes Convolution, Pooling, Concat, Dropout, Fully connected and Softmax. Convolution is Image transformation with the kernel on each pixel and its local neighbors on the whole image. The Inception V3 model was released in 2015, using 42 layers and a smaller error rate than its predecessors (Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, 2016).

As these “Inception modules” are stacked on top of each other, their output correlation statistics are bound to vary: as features of higher abstraction are captured by higher layers, their spatial concentration is expected to decrease suggesting that the ratio of $3 \times 3$ and $5 \times 5$ convolutions should increase as we move to higher layers (Szegedy, 2014).

One big problem with the above modules, at least in this naive form, is that even a modest number of $5 \times 5$ convolutions can be prohibitively expensive on top of a convolutional layer with a large number of filters. This problem becomes even more pronounced once pooling units are added to the mix: their number of output filters equals to the number of filters in the previous stage. The merging of the output of the pooling layer with the outputs of convolutional layers would lead to an inevitable increase in the number of outputs from stage to stage. Even while this architecture might cover the optimal sparse structure, it would do it very inefficiently, leading to a computational blow up within a few stages (Szegedy, 2014).

This leads to the second idea of the proposed architecture: judiciously applying dimension reductions and projections wherever the computational requirements would increase too much otherwise. This is based on the success of embeddings: even low dimensional embeddings might contain a lot of information about a relatively large image patch. However, embeddings represent information in a dense, compressed form and compressed information is harder to model. We would like to keep our representation sparse at most places (as required by the conditions of (Arora et al., 2014)) and compress the signals only whenever they have to be aggregated en masse. That is, $1 \times 1$ convolutions are used to compute reductions before the expensive $3 \times 3$ and $5 \times 5$ convolutions. Besides being used as reductions, they also include the use of rectified linear activation which makes them dual-purpose. The final result is depicted in Figure 5.

Table 3. Inception_V3 architecture

<table>
<thead>
<tr>
<th>Type</th>
<th>Patch size/stride</th>
<th>Input size</th>
</tr>
</thead>
<tbody>
<tr>
<td>conv</td>
<td>$3 \times 3$/$2$</td>
<td>$299 \times 299 \times 3$</td>
</tr>
<tr>
<td>conv</td>
<td>$3 \times 3$/$1$</td>
<td>$149 \times 149 \times 32$</td>
</tr>
<tr>
<td>conv padded</td>
<td>$3 \times 3$/$1$</td>
<td>$147 \times 147 \times 32$</td>
</tr>
<tr>
<td>pool</td>
<td>$3 \times 3$/$2$</td>
<td>$147 \times 147 \times 64$</td>
</tr>
<tr>
<td>conv</td>
<td>$3 \times 3$/$1$</td>
<td>$73 \times 73 \times 64$</td>
</tr>
<tr>
<td>conv</td>
<td>$3 \times 3$/$2$</td>
<td>$71 \times 71 \times 80$</td>
</tr>
<tr>
<td>conv</td>
<td>$3 \times 3$/$1$</td>
<td>$35 \times 35 \times 192$</td>
</tr>
<tr>
<td>$3 \times$ Inception</td>
<td>Inception at figure 5</td>
<td>$35 \times 35 \times 288$</td>
</tr>
<tr>
<td>$5 \times$ Inception</td>
<td>Inception at figure 5</td>
<td>$17 \times 17 \times 768$</td>
</tr>
<tr>
<td>$2 \times$ Inception</td>
<td>Inception at figure 5</td>
<td>$8 \times 8 \times 1280$</td>
</tr>
<tr>
<td>pool</td>
<td>$8 \times 8$</td>
<td>$8 \times 8 \times 2048$</td>
</tr>
<tr>
<td>linear</td>
<td>Logits</td>
<td>$1 \times 1 \times 2048$</td>
</tr>
<tr>
<td>softmax</td>
<td>classifier</td>
<td>$1 \times 1 \times 1000$</td>
</tr>
</tbody>
</table>

*name of corresponding author
Although VGGNet (Karen Simonyan, n.d.) has the compelling feature of architectural simplicity, this comes at a high cost: evaluating the network requires a lot of computation. On the other hand, the Inception architecture of GoogLeNet [20] was also designed to perform well even under strict constraints on memory and computational budget (Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, 2016).

**Dataset**
The dataset used in this processing is a dataset from food2k, where the dataset has been widely used as a public dataset for various research purposes on food. Deep learning in this study uses the dataset as shown in figure 6.

![Fig. 6 Food and drink image Dataset for Deep Learning](Source : google image)

**RESULT**
In this section the result is the appearance of the prediction that has been processed using the Convolutional Neural Network. Where is the use of the VGG19, ResNet50 and Inception_V3 architectures to classify the food2k dataset.

![Fig 7. The results of the detection/prediction of the food rating](Source : researcher property)

The accuracy of the Convolutional Neural Network model is as follows:

<table>
<thead>
<tr>
<th>Rating</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
</tr>
</tbody>
</table>

*name of corresponding author*
The loss of the VGG19 model is good, in the sense that it is very small, which is 0.0671. Accuracy is quite good where the results reached 96.86. The loss for the ResNet50 model has a value of 0.0718. Accuracy for Resnet reached 97.29. The loss of the Inception_V3 model has a value of 0.0686. Accuracy for Inception_V3 reached 97.57. The smallest loss is VGG19 and the best Accuracy is Inception_V3. So actually the three accuracy models are the best if the model is used to predict food. It can also be applied to several marketplace industries as an initial step in evaluating a food product.

From the description above, it can be discussed about the problems for providing a review or rating on the marketplace. The dataset used uses various public datasets. The dataset starts from the dataset type rating 1, rating 2, rating 3, rating 4 and rating 5. This solution can be used to find out or predict which food belongs to which rating. This is an answer from several marketplaces waiting for buyers to provide comments or ratings about the product. This rating can be used as an initial display of the assessment or review of a food product. By using a good appearance, color, position in getting pictures from came, it can be ascertained that the food product will be rated 5 or very good. Because of the problems in the collection of images of food products in the marketplace, it can be used as an initial review, so the results of this study are useful as an initial assessment of a food product.

CONCLUSION
Explanation of various methods for predicting the assessment or review of food products using Convolutional Neural Network with various architectures VGG19, ResNet50 and Inception_V3. All three models are able to predict food ratings. This is part of sentiment analysis with image input. So that sentiment analysis is not focused on using only text or sentences in making predictions.

SUGGESTION
Research can be continued by adding text analysis, to make further research. The use of images and text or sentences as input in the assessment for sentiment analysis. The use of pictures and sentences for further research.

REFERENCES


Yang, L., Li, Y., Wang, J., & Sherratt, R. S. (2020). Sentiment Analysis for E-Commerce Product Reviews in