Erasmus School of Economics

Master Thesis [Economics and Business Economics]

Building an image-based restaurant recommendation system

Name student: Lotte Pestman

Student ID number: 510775

Supervisor: A.C.D. Donkers

Second assessor: E. Raviv

Date final version: 11-07-2023

Contents

1 Introduction 3
1.1 Introduction to restaurant recommendation systems 3
1.2 Problem statement and central research question 3
1.3 Academic and managerial relevance 4
1.4 Summary of chapters 5
2 Literature 6
2.1 Collaborative filtering versus content-based recommender systems 6
2.2 The added value of images 7
2.3 The effect of images on consumer decision making 7
2.4 The effect of images on the perception of a restaurant 8
2.5 Image-based recommendations 9
2.6 Image labelling and tagging 10
3 Data 11
4 Methodology 12
4.1 Concepts model analysis 12
4.2 Word embeddings analysis 14
4.3 Image embeddings analysis 16
4.4 A dive into Neural Networks 17
5 Analysis and Results 19
5.1 Concepts model results 19
5.1.1 Analysis of the concepts in the model 19
5.1.2 Analysis of missing values and cosine similarities 22
5.1.3 Analysis of the recommendation results 23
5.2 Word embeddings model results 27
5.3 Image embeddings model results 29
5.4 Comparison of all models 32
5.5 Atmosphere versus food 34
6 Conclusion and discussion 36
References 38

1 Introduction

1.1 Introduction to restaurant recommendation systems

Nowadays every restaurant has their own website with all sorts of information about the restaurant, like photos of the restaurant and a menu. To make things easier for the consumer, platforms like TripAdvisor arose, which aggregate all restaurants on their website. This definitely makes it easier to find and compare restaurants, but with so many available restaurants, how does one decide on a restaurant? Here, the subject of this research, namely recommender systems, comes into play. A recommender system makes personalized recommendations to you based on your behaviour and preferences and/or the behaviour and preferences of others (Burke, 2002). It can do this by comparing you to other people who viewed or liked the same thing.

These restaurant review platforms have three sources of information per restaurant. First off, they have general information such as the price range, cuisine, location, and more. Secondly, they have the written reviews and ratings of past customers. Lastly, they show pictures that are uploaded by past customers or by management of the restaurant. Through all the information provided by past customers, you are able to get more insights into the atmosphere, the quality of the food, the customer service, and much more. These platforms allowed for more information exchange between customers through reviews. This became a core part of a recommendation system, as there is much more information in reviews than in just descriptions of restaurants (Al-Ghuribi & Noah, 2019). While much research has been done on the first two sources of information, research on the third is limited. In this thesis I will study the information content of images to create recommendations. For most websites, the majority of the users do not have an account. This is no different for restaurant review platforms and thus I will create a recommendation system without having data on the user, except for what they are currently viewing. Therefore, the recommendation system will make 'more like this' recommendations.

1.2 Problem statement and central research question

Recommender systems are still being further developed and improved. However, most recommender systems do not take images into account (Chu & Tsai, 2017). This is however a missed opportunity, as images contain much more information than text. "A picture is worth a thousand words" is especially true for a restaurant recommendation system. How the food looks, the atmosphere of the restaurant, and much more is difficult to describe, but can easily be captured in a picture. Thus, by not including images of the restaurant, you lose valuable and essential information. Using images will possibly make

the recommendation more accurate and that is why this research aims to create a restaurant recommendation system which considers all aspects of a restaurant by using image data. This will be done specifically for TripAdvisor as they are one of the biggest review platforms for restaurants, but they currently do not have a personalized recommendation system on their website. On their website the restaurants are sorted by rating and the users can filter restaurants based on locations, cuisine, price range, and such. When looking at a certain restaurant, TripAdvisor recommends the best rated restaurants in the same price category that are near that restaurant. As stated before, these recommendations might be improved by creating a system which relies on similarities between restaurants based on their images. This leads us to the research question:

What restaurant features are important in a restaurant recommendation system for TripAdvisor which uses image data of restaurants in the Netherlands?

1.3 Academic and managerial relevance

This research is academically relevant as it provides more information of recommendation systems and more specifically using image data in recommendation systems. As said before, most recommendation systems only use text data, even though there is valuable information in images. In the process of creating a recommendation system, this research will also dive further in image analysis. This is also academically relevant as images have already become an essential part of the way we communicate online, and this will only increase in the future, but there is currently only a limited amount of information available on this subject.

This research is managerially relevant in a broad sense. A recommendation system which can include images of the restaurant will possibly generate more accurate recommendations. TripAdvisor currently has a quite general recommendation system, which only takes general restaurant information and average rating into account. This research will show TripAdvisor and other restaurant review platforms the benefits of using an image-based restaurant recommendation system, as it is implementable and accurate for everyone using the website.

1.4 Summary of chapters

In the upcoming chapters, the introduced problem will be investigated. First, previous work will be researched to understand better how recommender systems are made, what information lies in images, and how images have been used in some type of recommendation system. Then, the data is described in chapter 3. In chapter 4, the methodology will be covered, giving a first glimpse of the models created in this research. Chapter 5 dives into the results of the model and gives an analysis of these obtained results. Chapter 6 closes this research with a conclusion based on the results and a discussion.

2 Literature

2.1 Collaborative filtering versus content-based recommender systems

In recommender systems there are two main types of algorithms used, namely collaborative filtering methods and content-based methods. A combination of these methods is called a hybrid recommender system (Aggarwal, 2016).

Collaborative filtering relies heavily on the information of all users and all items, as it compares all users to one another to find and match profiles which are similar to each other in their preferences and interests. It bases its recommendations on these matched profiles, as they are likely to enjoy the same type of items (Isinkaye, Folajimi & Ojokoh, 2015). This is called user-based collaborative filtering. Item-based collaborative filtering on the other hand focuses more on finding similarities between the items than on finding similarities between users. It does this by comparing the reviews of item pairs to each other for all users that wrote a review about both items (Sarwar, Karypis, Konstan & Riedl, 2001). Based on this the algorithm can compute a similarity measure to determine which item pairs are reviewed similarly and which are not. Thus, to create a recommender system that uses collaborative filtering, you must have much data on all users of the website and particularly on their preferences through their ratings on all items.

Content-based methods on the other hand, focus more on the data on attributes of a product/service to determine its recommendations. It can do this by comparing the features of items that a user has liked in the past to features of other items on the website to determine what they might also like (Isinkaye, Folajimi & Ojokoh, 2015). Thus, a content-based recommendation system does not need information on other users, it only needs to know what the single user, for whom the recommendation is intended, likes. To determine what a user likes, we can use two types of data, namely explicit and implicit feedback. Actively liking and rating a product is a form of explicit feedback, but to do so a user needs an account and many users of a website do not have an account. In that case we can use implicit feedback, which are actions like viewing a product as this indicates interest in the product (Lops, de Gemmis & Semeraro, 2011). When we have more information on the user through an account where they rated and reviewed multiple items (explicit feedback), the recommendations can be based on the similarities between the rated and reviewed items and the recommended item. On the other hand, implicit feedback can be used to create a ‘more like this’ recommendation system which recommends items that are similar to the one being viewed by the user (Coelho et al, 2023). This is particularly useful when little data on the user is available as it only compares the features of a single item, namely the item that is being viewed, to the features of other items that can be recommended.

2.2 The added value of images

People are taking more and more photos, even of everyday activities like eating (Diehl, Zauberman & Barasch, 2016). As a result, images are becoming a large part of online communication through platforms such as Instagram, where large quantities of photos are shared on a daily basis (Liu, Dzyabura & Mizik, 2020). Images are even replacing text as the preferred medium of online communication according to Liu et al. (2020). In the digital world we now live in, the attention span of people online is limited and therefore images are becoming more important as they are a way to communicate information fast and efficiently (Pittman & Reich, 2016).

An experience can not fully be captured by language and thus need images. Images can be more powerful than text. Images trigger more emotions than just conversation or text and also lead to higher levels of involvement, especially regarding experiences (Reavey, 2011). According to Sundar (2008), people have a heuristic that pictures are more credible and trustworthy than text. Important here is that people value user generated photos more than the photos that the restaurant provides (Oliveira and Casais, 2019). TripAdvisor is ideal for this as almost all pictures are uploaded by the users themselves.

It is also important to understand what drives the decision of a consumer in choosing a restaurant. There are several factors that play a role such as the quality of food and service and the environment. To get an impression of these factors, consumers actively search online for past experiences from others (Oliveira & Casais, 2019). Meyers-Levy and Zhu (2008) propose that architectural factors and free-standing indoor structures influence our shopping process and decision making. The environment, the architecture and to some degree the quality of food can easily be captured in an image, but less so in text. This again highlights the importance of the use of images in a restaurant recommendation system.

2.3 The effect of images on consumer decision making

The consumer decision making process is an important factor when studying the use of images in a recommendation system. The consumer decision making process can be split up into five steps. First, the consumer enters the process by recognizing a problem. Secondly, the consumer searches information which they then use in the third stage, namely the evaluation of alternatives. Here they actively compare different products/services to one another to determine what they want. They then continue to purchase the best fitting product/service in the fourth stage, and they conclude their process in the fifth stage, which is post-purchase evaluation (Engel, Blackwell, Miniard, 1995). For this

research, the most important stages are the second and third stage as people are actively searching for information about restaurants and comparing them before making a decision.

Emotions play an important role in the decision-making process. Restaurant reviews that trigger a more emotional reaction increase the intention to visit that particular restaurant compared to reviews that do not trigger any emotions (Ruiz-Mafe, Chatzipanagiotou & Curras-Perez, 2018). As discussed before, images trigger more emotions than text, thus including images may increase the intention to visit more than a only textual review. Pictures also foster more positive attitudes compared to text (Bigne, Chatzipanagiotou & Ruiz, 2020).

In behavioural psychology, the term negativity bias is widely used. This term is also relevant for the information search stage in the consumer decision making process. Negativity bias refers to the greater effect that negative events have than positive events, as negative events are perceived as more dominant (Rozin & Royzman, 2001). Bigne, Chatzipanagiotou and Ruiz (2020) studied this effect in restaurant reviews and discovered that including a positive picture as the end of a negative review sequence significantly decreased the negativity bias more than a positive text. They conclude in their research that consumers should always include images in their information search as they can then make more informed and less biased decisions.

In a study on e-commerce, Mou and Shin (2018) found through eye-tracking that people pay more attention to pictures rather than the textual information that was provided when shopping online. Sudha and Sheena (2017) also found that pictures greatly influence people to buy clothes as most people are attracted to the pictures of clothes rather than text or videos. Based on these researches, it seems that images are the biggest part of the information search stage.

2.4 The effect of images on the perception of a restaurant

When dining out, one may expect the food and the service to be the most important part of the experience, but according to Kotler (1973) the atmosphere of a place is just as important. The physical surroundings of a restaurant, such as the décor, spatial layout and the ambience, have a large effect on consumer behaviour through customer satisfaction (Han & Ryu, 2009). The physical surroundings also have an effect on price perception. Consumer perception of a reasonable price lies mainly in the perceived quality, where the physical context plays an important role (Han & Ryu, 2009). Dogru and Pekin (2017) researched Airbnb accommodations and found that a 1% increase in the number of photos increases the price of the accommodation by 1%. More photos of the listing seems to justify a higher price. Teubner, Hawlitschek, and Dann (2017) similarly conclude that more photos lead to a higher listing price as the host seems to be more credible to consumers. This theory might

also be applicable to restaurants, which means that restaurants with more photos would typically be perceived as more credible and might also ask higher prices.

The number of photos that are posted by consumers seem to be related to the success rate of a restaurant according to Zhang and Luo (2023). Consumers specifically value the more informative photos more than aesthetic factors of the photo such as brightness, as informativeness of a photo is a stronger predictor of the survival of a restaurant. Photos of the food correlate the highest with the survival of a restaurant and interior photos correlate the second highest. However, this does not mean that the aesthetic features of an image are irrelevant. Zhang, Lee, Singh and Srinivasan (2016) conclude in their research on Airbnb listings that consumers do value high quality photos as it leads to more interest and demand in an apartment compared to low quality photos of the same apartment. This in turn leads to a higher yearly income for the host. Specifically, the colour attributes of a photo, such as brightness seem to be important in increasing demand. Thus, both the quality of the photo as the informativeness of the photo influence the perception of a restaurant.

2.5 Image-based recommendations

With this increase in the relevance of images, researchers have also taken more of an interest in the use of images in recommender systems. For example, Ay, Aydın, Koyun, and Demir (2019) created an image-based recommendation system for e-commerce platforms. They created their own model which is based on Information Maximizing Generative Adversarial Networks (InfoGAN). GAN's are neural networks that are trained on unlabelled data. With the input of a single picture of a shoe, they were able to make recommendations of other shoes with an accuracy of 84%.

McAuley, Targett, Shi and Van Den Hengel (2015) have broadened the use of e-commerce recommendation systems by creating a system which recommends complementary products instead of a substitute. So, for an image of a T-shirt, their model is able to recommend matching jeans, shoes, and accessories, based on the style of the T-shirt. They obtain an accuracy of 90% for men's clothing and 88% for women's clothing. Similarly, Yu, Zhang, He, Chen, Xiong, and Qin (2018) developed their recommendation system based on the classification of a product and the aesthetics of a product. For the classification, they use a Convolutional Neural Network (CNN) which extracts the necessary features from an image to determine what type of product it is. For the aesthetics, they use a Brain-inspired Deep Network (BDN) which can extract features that are more related to aesthetics that a CNN can not extract. Their results show that including this aesthetics factor truly enhances model performance.

Chu and Tsai (2017) have created a recommendation system that is similar to the one in this research. However, a key difference is that they also include information on the users, which leads to a hybrid recommendation system. They incorporate images as well as text data in their restaurant recommendation system. They conclude that including images does lead to an increase in the performance of the recommendation system compared to a recommendation system that does not use images. He and McAuley (2016) similarly find that including images in a recommender system leads to more accurate results, but they also recognize that it relieves the dreaded cold start issue that many recommender systems suffer from. From all these examples, we can conclude that an image-based recommendation system can be very accurate and that adding images in a recommendation system leads to a higher accuracy than an only text-based recommendation system and thus an image-based recommendation system is extremely relevant.

2.6 Image labelling and tagging

To determine the recommendations, the system must know what is visible in the image. Text-mining is already very developed, but image-mining is just getting started, even though images are becoming more relevant every day (Liu, Dzyabura & Mizik, 2020). The difference between labels and tags is that labels are a more high-level and general classification, while tags describe certain attributes in a picture (Wang et al., 2017). There are multiple ways to label or tag an image. In section 2.3 CNN was already mentioned as a way to classify the image as a certain product. Liu, et al. (2020) also used a multi-label CNN to determine whether certain perceptual attributes are present in an image. They apply this in the context of brand-related images on social media and achieved an accuracy of 90%. They trained their own CNN, but there are also already pre-trained models available that can label/tag images. For example, Dzyabura and Peres (2021) used the image tagging tool from Clarifai, which is again a deep CNN, to obtain tags for photos that are related to certain brands and subsequently used LDA topic modelling with the obtained tags to determine brand perception.

3 Data

At first 333 restaurants were obtained by web-scraping the restaurants in the province South-Holland from TripAdvisor which fall into 3 different cuisine types and 2 different price ranges. Asian, French, and Italian cuisines and cheap and expensive price ranges were selected. The first 10 images of the restaurants were scraped by a self-made web-scraper and the images were stored in a separate folder per restaurant and these folders were separated based on the cuisine and price category combination of the restaurants. The images that were scraped could either by uploaded by users that visited the restaurant or by the management of the restaurant. After removing all restaurants which fell into 2 or more cuisine categories and restaurants with 5 or less pictures, 240 restaurants were left with 6 or more pictures per restaurant. The exact number of restaurants per cuisine and price category can be found in table 1.

Table 1: Distribution of restaurants over the different cuisine and price categories

	Asian	**French**	**Italian**	***Total***
Cheap	99	14	56	*169 (70%)*
Expensive	12	56	3	*71 (30%)*
Total	*111 (46%)*	*70 (29%)*	*59 (25%)*	*240*

4 Methodology

4.1 Concepts model analysis

I first start out by making a simple model as a proof of concept. For this model, the idea is to compare the lists of atmosphere concepts and food concepts that are recognized in images of a restaurant to all other restaurants in the dataset and recommend the restaurant which is most similar in the concepts that are mentioned. To obtain the concepts, the images are analysed by using pre-trained models for image classification, namely the general image recognition vit model and the food item v1 recognition model from Clarifai. The former model can recognize concepts such as objects, themes, and moods in images. The latter model is more specific and can recognize dishes and ingredients from pictures and can return these concepts that are visible in the pictures. Evidently, the general model is much broader than the food model. To obtain the most accurate results, there is a need to analyse images of the interior, exterior, and such through the general model and analyse the images of the food and drinks through the food model. Therefore, it is necessary to separate the images based on their content. Each image of a restaurant is first analysed by the general model, which gives 10 concepts it recognizes in the image. If the word 'food' or the word 'drink' is one of the recognized concepts, the image is classified as a food image. If those words are not mentioned, the image is classified as an atmosphere image. For the atmosphere images, the concepts produced by the general images are added to a list of 'atmosphere concepts' of the restaurant in question. The food images are analysed by the food model which again produces 10 concepts. These food concepts are added to a list of 'food concepts' of the restaurant in question. However, pictures of food can also contain an atmosphere component, such as how the dish is presented, or the photo can contain both food and the interior/exterior of the restaurant. To account for the atmosphere component in the food pictures, the concepts that are produced by the general model for the food pictures are also added in a third variable called food atmosphere. Concepts that can be recognized by the food model are removed from the food atmosphere concepts, to ensure that the food atmosphere variable only contains atmosphere concepts. The word 'food' is not mentioned in the food concepts as it goes without saying that the pictures contain food. However, this word is mentioned by the general model and thus mentioned many times in the food atmosphere concepts, but it adds very little value as it is per definition visible in all food atmosphere photos. The high frequency of 'food' leads to high cosine similarities which puts much emphasis on the similarity in the number of pictures that contain food. Therefore, the word 'food' is also deleted from the food atmosphere concepts.

To compare the lists of concepts per restaurant, a similarity measure between two lists of words is needed. This research will consist of lists of words and embeddings. To compare the performance between all models in this research, there is a need for a consistent similarity measure. The other

models in this research will deal with embeddings of words and images, and as a list of terms can also be seen as a vector, the cosine similarity is chosen as the similarity measure. It is typically chosen as the similarity measure for vectors (Singhal, 2001). The angle between two vectors is used as a divergence measure and the cosine of this angle will then be the numerical representation of the similarity between vectors. Another similarity measure is the dot product of two vectors. However, the magnitudes of the vectors are taken into account in the dot product, while the direction of the vector is much more important in determining the similarity between restaurants than the length of the vector. The Euclidean distance is also often used as a similarity measure, however this measure is also sensitive to the magnitudes of vectors and is thus again not a desirable measure (Xia, Zhang & Li, 2015).

Thus, the cosine similarity was chosen for this model, which will compute how similar two lists are by comparing word occurrences in the two documents or lists. The cosine similarity can be measured by the formula $\cos(\theta) = \frac{A \times B}{||A||\,||B||}$, where A x B is the dot product and ||A|| and ||B|| are the magnitudes of A and B respectively (Li & Han, 2013). In the case of lists of words, A and B are the lists of words for either the atmosphere concepts or the food concepts of the queried restaurant and another restaurant. First, all concepts that are mentioned in the two lists, A and B, are gathered in a terms list. Then the dot product is calculated by summing up the multiplication of the number of times a term is mentioned in list A and the number of times it is mentioned in list B for every term in the terms list. The magnitude of list A is calculated by summing up the squares of the number of times a term is mentioned in list A for every term in the terms list and taking the square root of the summation. The magnitude of list B is calculated similarly. The number of times the terms are mentioned can also be seen as creating a new vector. If the i-th term in the terms list is not mentioned in list A, the vector of A will have a 0 at the i-th place. If the i-th term is mentioned once in list A, the vector of A will have a 1 at the i-th place. Then, the cosine similarity formula can be rewritten as $\cos(\theta) = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$, where A_i and B_i are the i-th component of the vectors of A and B respectively and n is the number of components in a vector, which would be equal to the number of terms in the terms list (Li & Han, 2013). The cosine similarity ranges from -1 to 1 (Nguyen & Bai, 2010). However, in this model the term frequency is used which can not be negative. Thus, the cosine similarity will never be below zero in this model. If A and B are exactly the same, the two vectors are pointed in exactly the same direction, which makes the angle between them 0 degrees and thus the cosine similarity will be equal to one. All terms in list A will also be in list B, therefore the dot product of A and B will be equal to the multiplication of the magnitudes of A and B. If A and B do not have any terms in common, the cosine similarity will be equal to zero. The cosine similarity was measured for both the atmosphere concepts and the food concepts separately. Thus, each individual restaurant has two cosine similarities with all other restaurants in the dataset, namely

the similarity between the atmosphere photos and the similarity between the food photos. To obtain a single cosine similarity value, the average between the atmosphere similarity and the food similarity was taken.

Now, recommendations can be made based on the images of the restaurants, or more specifically based on the concepts that are recognized in those images. By comparing the recognized concepts to each other through the cosine similarity measurement, we check whether the same concepts are recognized in the photos of the restaurants. If the restaurants serve the same dishes or use the same ingredients, they are more alike. These dishes and ingredients will be recognized in images by the classification model and will be mentioned in the concepts of the restaurants. These concepts will thus be mostly the same and this will result in a high cosine similarity. Likewise, if they style their restaurant the same way, the same concepts will come up in the concepts of the atmosphere photos and thus they will be more alike. The higher the cosine similarity, the more alike the two restaurants are. After all cosine similarities are calculated for a query restaurant, the restaurant in the dataset with the highest average cosine similarity is recommended.

4.2 Word embeddings analysis

The word occurrence of the concepts model only recognizes whether words are exactly the same or not. Thus, the words ‘spaghetti’ and ‘lasagna’ are just as different as the words ‘spaghetti’ and ‘sushi’, while in fact lasagna is much more similar to spaghetti than sushi is to spaghetti. It can also be said that if you enjoy spaghetti, you are more likely to enjoy lasagna than sushi. Therefore, the recommendation should take the meaning and relationships of words into account. To bring more nuance in the way the similarity is measured, word embeddings might be the way to go. Word embeddings are vectors in a multi-dimensional space which describe the meaning of words. The vectors of words with a similar meaning will be closer together than vectors of words with a very different meaning (Yu, Wang, Lai & Zhang, 2017). Each word can get its own embedding that exactly represents that word. To build upon the previous example, the vector of the word ‘spaghetti’ will be closer to the vector of the word ‘lasagna’ than to the vector of the word ‘sushi’, as can be seen in a simple two-dimensional example in figure 1. The angle b is smaller than angle a, meaning that the cosine similarity cos(b) is larger than the cosine similarity cos(a). Now it can be concluded that lasagna and spaghetti are more similar to one another than sushi and spaghetti. This is a simple example of words, but in this research, there are lists of words that need to be compared and a single vector to describe the whole list is more efficient than multiple vectors per list.

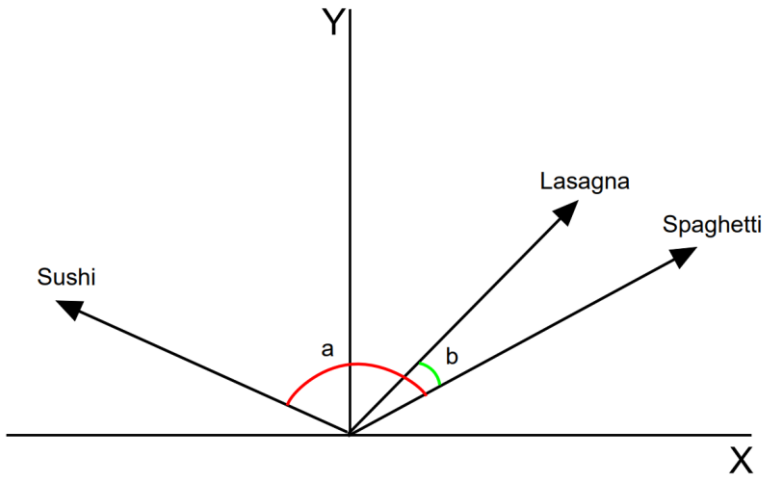

Figure 1: Example two-dimensional word embedding vectors

SpaCy is a Natural Language Processing (NLP) library in Python, which has built-in word embeddings. The en_core_web_md pre-trained pipeline has 20.000 unique 300-dimensional vectors. First, spaCy tokenizes a document. It splits up the document into words and punctuation, which are now called tokens. Each token can be transformed into a built-in vector that describes the meaning of the token. When applied on a document with multiple tokens, the function will take the average of all token vectors in de document. Thus, through this library it is possible to obtain two vectors per restaurant that describe the two lists of concepts of the atmosphere and the food of a restaurant.

The cosine similarity was taken as the similarity measure between vectors. The calculation of this measurement is now slightly different compared to the last model. Where a vector was created based on the term frequencies in the previous model, now there is already an existing vector. The formula of course stays the same, namely $\cos(\theta) = \frac{A \times B}{||A||\,||B||} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$, where A_i and B_i are the i-th component of the vectors of A and B respectively and n is the number of components in a vector, which will be 300 numbers in this case. Contrary to the previous model, the cosine similarity can now also be negative because the components of a vector are sometimes negative. Again, the similarity will be close to 1 if the vectors are near each other. The similarity will be equal to 0 if the vectors are orthogonal and it will be -1 if the vectors are exact opposites (Liu, 2014). The cosine similarity was calculated between the atmosphere vectors, food vectors, and food atmosphere vectors separately, and the average of the three similarity values was taken to obtain a single cosine similarity between two restaurants.

The restaurant with the highest cosine similarity would be most similar in atmosphere and food to the query restaurant, as the vector portrayal of the concepts recognized in the pictures are closest together and thus similar to one another. They would have similar dishes, but they would not necessarily be exactly the same. Contrary to the previous model, an Italian cheap restaurant with only pictures of spaghetti can now also be recommended to an Italian cheap restaurant with pictures of lasagna. Again, the restaurant with the highest cosine similarity is recommended based on the query restaurant.

4.3 Image embeddings analysis

The embeddings can also be taken a step further. Instead of analysing which concepts are recognized in the images and transforming these concepts into embeddings, it is also possible to convert the entire image into an embedding. Then, how visually similar the images are of restaurants will be measured instead of how similar the concepts in those images are. This may be a more direct way of measuring similarity and may also be able to capture more information available in the picture. Thus, the atmosphere and food atmosphere images are embedded using the general image embedding vit from Clarifai, which returns a 768-dimensional vector of an image. The food images are embedded by the food item v1 image embedding from Clarifai, which embeds the image into a 1024-dimensional vector. These image embedders are often a Convolutional Neural Network (CNN) due to their high performance (Bell & Bala, 2015). Clarifai wrote in their blog post on August 4, 2017, that they use CNNs for their visual recognition. Based on this limited information that is available, it was concluded that Clarifai also uses CNNs for their visual classifiers and embedders. The working of a CNN will be explained in the next section.

With these image embeddings, the cosine similarity can again be calculated between restaurants in the same way as in the word embeddings model. The average cosine similarity between all atmosphere photos of the query restaurant and all atmosphere photos of another restaurant is taken as the cosine similarity for atmosphere. Thus, if the query restaurant has 4 atmosphere photos and the test restaurant has 5 atmosphere photos, there will be 4 x 5 = 20 cosine similarities, which will be averaged into a single cosine similarity. The same is done to calculate the cosine similarity for food and the cosine similarity for food atmosphere. Then, the average of the three similarity values is taken to obtain a final cosine similarity value. The recommendation is done based on the final values by selecting the highest value.

Now the images that are visually most similar to one another will have the highest cosine similarity value. Thus, the restaurant where all images are on average visually the most similar to the query restaurant's images is recommended.

4.4 A dive into Neural Networks

Convolutional Neural Networks (CNNs) are often used in image recognition tasks. It does so by recognizing patterns in the images (Hijazi, Kumar & Rowen, 2015). CNNs are a type of Neural network (NN). NNs consist of multiple layers which are connected through their nodes or 'neurons'. The basic structure of a NN begins with an input layer, then one or multiple hidden layers and it ends with an output layer. The fully-connected hidden layers are responsible for adjusting the input to determine what the output should be, by learning from the previous layers (O'Shea & Nash, 2015). A CNN is unique due to its convolutional layers which work as a feature extractor. The input of one of the nodes in a convolutional layer is accompanied by the local neighbourhood around the input. The nodes then filter through the image in a grid-like manner (Albawi, Bayat, Al-Azawi, & Ucan, 2017). The element-by-element matrix multiplication and summation of these filters are weighted before passing it along as input for the next layer. A convolutional layer is always followed by a pooling layer which serves as a feature resolution reduction layer. This protects the feature against noise and manages the computational time (Hijazi, Kumar & Rowen, 2015). The output of convolutional and pooling layers is called a feature map and can be seen as a numerical representation of the features in the image (Albawi, Bayat, Al-Azawi, & Ucan, 2017). This multi-dimensional feature map can be flattened into a one-dimensional vector through a flattening layer in the CNN (Wang et al, 2020). This is the embedding that represents the image. Depending on the assigned task of the CNN, the model can stop here, or the (flattened) feature map can be fed through to fully-connected layers as in a standard NN. The fully-connected layers are able to correctly classify the image and/or the image features. For the image embeddings in section 4.3, the model is ended when the feature map is flattened by the flattening layer in the CNN. The output is then the flattened vector. For the concept in section 4.1, the feature map goes through the fully connected layer where classification takes place. In this case, the classification is a multi-label classification as multiple features in the image are classified and not a single-label classification, which classifies the whole image with a single label. The concepts are then the output of the CNN. CNNs can be trained for a variety of tasks. These tasks include tasks that are unrelated to images. This can be done through a one-dimensional CNN, which is created with sequential data, such as text and time series data. However, two-dimensional CNNs are used for the grid-like structure that images have. The tasks that fall under this type of CNN are image classification, object detection, image

segmentation and face recognition (Li et al, 2022). CNNs can also be created with 3- or more dimensional data, but this becomes hard to grasp for humans and is therefore rarely done.

Image recognition models are trained on a large number of labelled images (Hijazi, Kumar & Rowen, 2015). Because the images are labelled, the model learns which images and image features are similar to each other and which are different. Thus, the outcome of a CNN is dependent on the data it was trained on and the task it was trained for (Wu, He, Sun & Tan, 2018).

5 Analysis and Results

5.1 Concepts model results

5.1.1 Analysis of the concepts in the model

As described in the methodology section, the concepts were obtained through two pre-trained models from Clarifai. To get a feeling of which concepts are mentioned, an overview of the top 20 most frequent concepts with their frequencies are visible in table 2. Most atmosphere pictures are of the interior of the restaurant and show basic furniture such as tables and chairs. These concepts give an impression on what type of restaurant it is, for instance a bar, a shop or a true restaurant. The most frequent concepts of food are mainly the type of protein that is visible in the picture. They often start off broader with words like 'fish' and 'seafood', before mentioning more specific words such as 'salmon', and 'shrimp'. The food atmosphere concepts add more information on the type of meal it is, breakfast, lunch or dinner, and more descriptive words such as delicious and homemade.

Table 2: Top 20 most frequent concepts for atmosphere, food, and food atmosphere

Atmosphere		**Food**		**Food atmosphere**	
Concept	**Frequency**	**Concept**	**Frequency**	**Concept**	**Frequency**
Indoors	313	Chicken	598	Dinner	612
Restaurant	282	Sauce	525	Plate	530
Table	281	Vegetable	453	Lunch	505
Furniture	238	Cheese	426	No person	420
Stock	198	Pork	420	Delicious	364
People	168	Meat	402	Meal	363
Chair	165	Beef	347	Restaurant	319
Hotel	157	Rice	318	Epicure	258
City	156	Salad	300	Refreshment	224
No person	154	Sweet	285	Dish	190
Luxury	147	Cream	255	Drink	138
Shop	136	Fish	239	Table	133
Interior design	134	Chocolate	236	Cooking	123
Street	134	Cake	201	Appetizer	108
Outdoors	134	Seafood	187	Bowl	91
Bar	131	Tea	177	Homemade	91
Adult	124	Pepper	173	Sugar	79
Dining	123	Shrimp	168	Breakfast	76
Window	112	Salmon	165	Slice	74
Seat	103	Beer	153	Indoors	72

Then, in table 3 are the top 10 most frequent concepts for cheap restaurants and expensive restaurants. Logically, luxury is mentioned much more for expensive restaurants while shop and street are mentioned more for the cheap restaurants. Apparently, cheap restaurants have more people in their pictures, while for expensive restaurants there are typically no people in the picture. The only

notable difference for the food concepts is that rice is mentioned often for cheap restaurants while cream is more often mentioned for expensive restaurants. Epicure is more prominent in expensive restaurants and also the word 'plate' as the food is often presented on nice plates which are more prominently visible. Cheap restaurants serve lunch more often than expensive restaurants. To amplify the contrast between the cheap and expensive concepts, some of the unique concepts per category are presented in table 4. Unique concepts are concepts that are mentioned solely in that particular category. For example, the unique concepts of Italian are obtained by removing all concepts that are mentioned in French and Asian restaurants from the Italian concepts. Therefore, cheese is not mentioned in the unique concepts of Italian as pictures of French and/or Asian restaurants also contain cheese. Thus, the unique concepts contain concepts that are solely recognized in images of that particular category. Here, the difference becomes very obvious. Where cheap restaurants are simpler with the words 'diner', 'crust', and 'homemade', the expensive restaurants are more sophisticated with 'silverware', 'scallop', and 'indulgence'.

Table 3: Top 10 concepts for atmosphere, food, and food atmosphere per price category

Atmosphere concepts		**Food concepts**		**Food atmosphere concepts**	
Cheap	**Expensive**	**Cheap**	**Expensive**	**Cheap**	**Expensive**
Indoors	Indoors	Chicken	Chicken	Lunch	Plate
Restaurant	Table	Sauce	Sauce	Dinner	Dinner
Table	Restaurant	Pork	Vegetable	Plate	Epicure
Stock	Furniture	Vegetable	Cheese	No person	No person
Furniture	Hotel	Cheese	Meat	Meal	Delicious
People	Luxury	Meat	Pork	Restaurant	Lunch
City	Chair	Beef	Salad	Delicious	Refreshment
Chair	Stock	Rice	Beef	Refreshment	Meal
Shop	No person	Sweet	Sweet	Epicure	Restaurant
Street	Dining	Salad	Cream	Dish	Appetizer

Table 4: Unique concepts for atmosphere, food, and food atmosphere per price category

Atmosphere concepts		**Food concepts**		**Food atmosphere concepts**	
Cheap	**Expensive**	**Cheap**	**Expensive**	**Cheap**	**Expensive**
Design	Dishware	Crust	Scallop	Meal	Slice
Diner	Silverware	Cappuccino	Sea bass	Dish	Bowl
Vector	Tablecloth	Casserole	Halibut	Drink	Indulgence
Letter	Cutlery	Spring rolls	Oyster	Homemade	Dining
Tourist	Flatware	Fried rice	Panna cotta	Stock	Wood
Bike	Reception	Cheddar	Bass	Indoors	Helping
Pattern	Blooming	Cabbage	Radish	Shop	Healthy
Recreation	White wine	Kale	Tuna tartare	Breakfast	Creamy
Lantern	Napkin	Mocha	Rose	Hot	Vanilla
Pizzeria	Knife	Paella	Coulis	People	Luxury

Table 5: Top 10 concepts for atmosphere, food, and food atmosphere per cuisine category

Atmosphere concepts			Food concepts			Food atmosphere concepts		
Asian	**French**	**Italian**	**Asian**	**French**	**Italian**	**Asian**	**French**	**Italian**
Indoors	Indoors	Indoors	Chicken	Chicken	Chicken	Dinner	Plate	Lunch
Restaurant	Table	Restaurant	Sauce	Sauce	Sauce	Lunch	Dinner	Dinner
Table	Restaurant	Table	Pork	Vegetable	Cheese	Plate	Delicious	Delicious
Stock	Furniture	Furniture	Vegetable	Cheese	Vegetable	Meal	Epicure	No person
Furniture	Chair	Stock	Meat	Meat	Meat	No person	No person	Plate
People	Hotel	People	Beef	Pork	Pork	Restaurant	Refreshment	Restaurant
City	Luxury	City	Rice	Salad	Beef	Delicious	Lunch	Refreshment
Shop	Stock	No person	Cheese	Beef	Salad	Dish	Dish	Meal
Chair	No person	Chair	Salad	Sweet	Rice	Epicure	Restaurant	Epicure
Street	People	Hotel	Sweet	Cream	Sweet	Refreshment	Meal	Drink

Table 6: Unique concepts for atmosphere, food, and food atmosphere per cuisine category

Atmosphere concepts			Food concepts			Food atmosphere concepts		
Asian	**French**	**Italian**	**Asian**	**French**	**Italian**	**Asian**	**French**	**Italian**
Exhibition	Silverware	Pizzeria	Teriyaki	Puree	Crust	Shop	Bowl	Cup
Sun	Tablecloth	Absence	Spring rolls	Halibut	Pepperoni	Market	Healthy	Caffeine
Light	Cutlery	Fireplace	Fried rice	Panna cotta	Penne	People	Dining	Dawn
Competition	Garden	Refrigerator	Bibimbap	Blackberry	Ricotta	Tray	Creamy	Mug
Number	Shape	Caffeine	Chicken wings	Radish	Cocoa	Adult	Wood	Pub
Letter	Flatware	Coffee cup	Kebab	Tuna tartare	Fudge	Shopping	Nutrition	Indoors
Antique	Reception	Little	Takoyaki	Rose	Rye	Poultry	Bakery	Pizzeria
Chili	Park	Smile	Soy sauce	Macaron	Bruschetta	Counter	Luxury	Vanilla
Rack	Blank	Home appliance	Stir-fry	Mackerel	Quiche	Man	Wineglass	Cornet
Crowd	Show	Cooker	Rib	Coulis	Brownie	Spoon	Helping	Merchandise

Zooming in further, the concepts per cuisine are also compared in tables 5 and 6. First for the top 10 concepts per cuisine, Asian restaurants are more often a shop-like restaurant, while French restaurants have a more hotel-like luxury atmosphere. The Asian cuisine also has a lot of rice, while the French cuisine does not. The concept cheese is the highest in the top 10 for the Italian cuisine, likely due to the amount of cheese on pizzas and pastas. The plate seems to be the more important in the French cuisine, most likely due to the fact that most expensive restaurants are French. Therefore, the word epicure is also much higher for the French cuisine. Then for the unique concepts, the French cuisine is again more sophisticated with 'silverware' and 'tablecloth', while the atmosphere in Asian restaurants seems to be more related to sun and antique. The Italian restaurants seem to be a little homier with words like 'fireplace', 'smile' and 'refrigerator'. Typical dishes and ingredients are mentioned for all cuisines in the food concepts. In the food atmosphere concepts, the Italian cuisine concepts clearly represent the Italian coffee shops and ice cream shops with words like 'cup', 'caffeine', and 'cornet'. Some French restaurants are uniquely described as a bakery, likely referring to a typical French boulangerie, and the Asian restaurants are uniquely described with shop-like terms such as 'market', 'shopping', and 'counter'. The French cuisine also seems to be healthier than the Asian and Italian cuisine.

5.1.2 Analysis of missing values and cosine similarities

After further investigation of the concepts, it becomes apparent that there are 25 restaurants that do not have either atmosphere pictures or food pictures and thus there are some missing values for the concepts. The specific distribution of the type of photos available of the different cuisine and price categories can be seen in table 7. If the cosine similarity is measured between the atmosphere concepts of a query restaurant and a restaurant with a missing value for atmosphere, the cosine similarity will be equal to zero. If the average is then taken of the atmosphere, food, and food atmosphere similarities, the restaurants with a missing photo category will never be recommended as the average will always be low. Therefore, the average needs to be adjusted for these missing values by only including the cosine similarities of photo categories that are actually available. Thus, if there are no atmosphere photos available, then the average cosine similarity will consist of the food and food atmosphere cosine similarities. If there are no food photos available, the average cosine similarity will only consist of the atmosphere cosine similarity.

Table 7: Distribution of type of photos available

	Category	Both atmosphere and food photos	Only atmosphere photos	Only food photos
Price	Cheap	151	9	9
	Expensive	64	3	4
Cuisine	Asian	101	3	7
	French	63	2	5
	Italian	51	7	1
Cuisine and price	Asian cheap	89	3	7
	Asian expensive	12	0	0
	French cheap	13	0	1
	French expensive	50	2	4
	Italian cheap	49	6	1
	Italian expensive	2	1	0

The mean cosine similarity and the standard deviation excluding missing values are 0.178 and 0.169 for atmosphere, 0.322 and 0.166 for food, and 0.361 and 0.195 for food atmosphere respectively. Thus, food and food atmosphere similarity is much higher than the atmosphere similarity. The restaurants without food photos are then at a disadvantage as their average cosine similarity will consist only of their atmosphere similarity, which is typically lower than food similarity. To solve this problem, the cosine similarities are standardized per category before they are averaged across the available categories.

5.1.3 Analysis of the recommendation results

To assess the accuracy of the model, the number of times the model recommends a restaurant with the same cuisine and/or price category as the queried restaurant is counted and divided by the total number of restaurants with that cuisine and/or price category. The accuracies of all tested versions of the concepts model are visible in table 8. The expected accuracy for a random sample is also given as a benchmark to illustrate what the performance would be when no model is applied when giving recommendations. The recommendations are then randomly selected. If the model performs better than this benchmark, it can be said that the model adds value. The accuracy is assessed for the average cosine similarity of both the non-standardized values and the standardized values of atmosphere, food, and food atmosphere cosine similarities to compare the effect of standardization. Thus, for restaurants with an Asian cuisine, the non-standardized model recommends a restaurant with an Asian cuisine 82% of the time, while the standardized model recommends a restaurant with an Asian cuisine 79% of the time. In general, the standardized model slightly underperforms compared to the non-standardized model. The standardization using the mean and standard deviation might be suffering from a relative shift in weight, where the variable (atmosphere, food, or food atmosphere) with the least variance gets

relatively more weight. However, in the non-standardized version, restaurants with only atmosphere pictures are at a large, unfair disadvantage due to the low average cosine similarity of atmosphere photos compared to food photos. Therefore, the standardized version of the concepts model is still preferred, as it is the proper way of dealing with the difference in means between atmosphere and food and food atmosphere, and the difference in accuracy is negligible.

On average, the standardized model correctly recommends a restaurant with the same cuisine 71% of the time. The standardized model correctly recommends the same price category 79% of the time. The model correctly recommends the combination of cuisine and price category 63% of the time, but here, the different combinations have very different accuracies. The model never correctly recommends a restaurant for the Italian expensive category, however this is a very small sample with only three restaurants. The model seems to be performing better in categories where there are more restaurants like the Asian cheap restaurants, where there are 99 restaurants in the dataset, as it should be as the expected accuracy is higher for categories with relatively more restaurants. For the Asian cheap category, the accuracy based on random selection is at least 41%, but the standardized concepts model recommends an Asian cheap restaurant 76% of the time, which is much better.

Table 8: Accuracies of the non-standardized and standardized concepts model

	Category	**Accuracy random sample**	**All concepts without standardization**	**All concepts with standardization**
Price	Cheap	70%	79%	79%
	Expensive	30%	80%	79%
	Weighted average	*58%*	*80%*	*79%*
Cuisine	Asian	46%	82%	79%
	French	29%	71%	70%
	Italian	25%	56%	56%
	Weighted average	*36%*	*73%*	*71%*
Cuisine and price	Asian cheap	41%	77%	76%
	Asian expensive	5%	58%	33%
	French cheap	6%	0%	0%
	French expensive	23%	75%	73%
	Italian cheap	23%	55%	55%
	Italian expensive	1%	0%	0%
	Weighted average	*29%*	*65%*	*63%*

It is also interesting to look at actual pictures of the correct and incorrect recommendations the model makes to better understand how the model works. In figure 2 and 3, correct recommendations of the Asian cheap category are visible. In figure 2 the idea of multiple dishes as a sort of 'shared dining' idea was correctly detected by the model and also the type of food, for example some type of satay sticks, was correctly recognized. In figure 3, the query restaurant on the left is more of a shop-like

restaurant for take-away. This was again correctly recognized by the model and therefore the recommendation is also a shop-like restaurant that makes similar foods. For the Italian cheap restaurant Sugo, the model correctly recommends the same restaurant chain 'Sugo' but a different location, as these restaurant chains always have the same interior design and the same foods/drinks. The same goes for the restaurant chain 'Doppio espresso'.

Figure 2: Restaurant Snack Inn (left) and its correct recommendation Kopi Kopi (right)

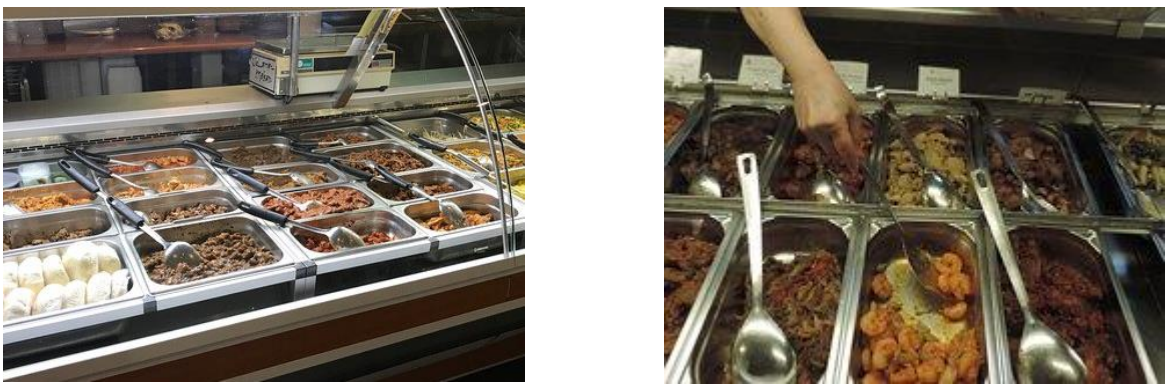

Figure 3: Tjendrawasih (left) and its correct recommendation Toko Dian (right)

On the other hand, in figure 4 the model fails by recommending a French expensive restaurant to a cheap Italian query restaurant. Both the food and the atmosphere are very different. When comparing the concepts of both restaurants to one another, it becomes evident that the model may be too simplistic. It recognizes furniture, chairs and tables, but can not recognize a difference in the chairs and tables. It also fails to notice the tablecloths, silverware and wine glasses on the table of Het

Prentenkabinet (French expensive). It also adds the concept ‘luxury’ to the atmosphere photos of Belicio Cheatday Dordrecht (Italian cheap), maybe due to the fact that it is white and clean. The recommendation must be done mainly on the basis of atmosphere similarity as the food concepts for Belicio Cheatday Dordrecht are mainly sweet while for Het Prentenkabinet they are mainly savoury. The food atmosphere similarity also lacks as Het Prentenkabinet has ‘delicious’ and ‘refreshment’ mentioned many times, while Belicio Cheatday Dordrecht has the words ‘indulgence’ and ‘buffet’ multiple times. It becomes evident from figure 5 that the model may not be entirely capable of recognizing the difference in quality of food. The Asian cheap iYumi has the Asian expensive Sushi Morikawa as a recommendation. They both serve sushi, but Sushi Morikawa has better quality fish. This difference is however not recognized in the concepts.

Figure 4: Belicio Cheatday Dordrecht (left) and its incorrect recommendation Het Prentenkabinet (right)

Figure 5: iYumi (left) and its incorrect recommendation Sushi Morikawa (right)

5.2 Word embeddings model results

The mean cosine similarity and the standard deviation excluding missing values are 0.686 and 0.141 for atmosphere, 0.907 and 0.052 for food, and 0.803 and 0.102 for food atmosphere respectively. The mean cosine similarities lie much higher than in the concepts model. Still, the similarities for food and food atmosphere are higher than the atmosphere similarity. The restaurants without food photos are then at a disadvantage as their average cosine similarity will consist only of their atmosphere similarity, which is typically lower than food similarity. Thus, the cosine similarities are standardized again.

The accuracies of the non-standardized and the standardized word embeddings model are visible in table 9. The accuracies are now slightly higher on average for the standardized version. Therefore, the preference for the standardized model is now even higher as it is the more proper way of dealing with the difference in means and the model performs better. The model is better at recommending an expensive restaurant correctly than a cheap restaurant. On average, the model correctly recommends a restaurant in the same price category 75% of the time. For the cuisine category, the accuracies vary, where the model is less accurate in recommendations for Italian restaurants than for French and Asian restaurants. The model never makes a correct recommendation for the French cheap and Italian expensive categories. However, the model does perform much better than the expected accuracy for the Asian cheap and French expensive categories. On average, the model correctly recommends a restaurant in the same cuisine and price category 54% of the time.

Table 9: Accuracies of the non-standardized and standardized word embeddings model

	Category	Accuracy random sample	All concepts without standardization	All concepts with standardization
Price	Cheap	70%	73%	74%
	Expensive	30%	79%	79%
	Weighted average	*58%*	*75%*	*75%*
Cuisine	Asian	46%	70%	71%
	French	29%	64%	66%
	Italian	25%	42%	44%
	Weighted average	*36%*	*62%*	*63%*
Cuisine and price	Asian cheap	41%	67%	69%
	Asian expensive	5%	42%	33%
	French cheap	6%	0%	0%
	French expensive	23%	63%	63%
	Italian cheap	23%	38%	39%
	Italian expensive	1%	0%	0%
	Weighted average	*29%*	*53%*	*54%*

However, accuracies do not capture the whole performance of the model. Which exact recommendations are made is also of interest to see how the model analyses the pictures. The concepts model has for the query restaurant Very Italian Pizza (see figure 6) the recommendation Happy Italy (see figure 7). Both these restaurants are Italian cheap restaurants. Here the similarity is obvious, with many pictures of either pizza or pasta. However, Very Italian Pizza did have more pictures of pizza than of pasta, while for Happy Italy's pictures the ratio was more balanced. Now for the word embeddings model the recommendation for Very Italian Pizza changed to the restaurant La Bella Rosa (see figure 8). Even though this is also an Italian cheap restaurant, the similarity becomes less obvious. La Bella Rosa does not only have pictures pizza and pasta, but it also includes other typical Italian dishes and ingredients, such as risotto, bruschetta and gambas. This is a perfect example of the added value of transforming concepts into word embeddings. The recommendations can be broader than the recommendations of the concepts model as restaurants with other concepts than the query restaurant's concepts are not immediately seen as 'wrong'. Rather, the word embeddings can be closer together or further apart. This allows for risotto and bruschetta to also be similar to pizza and pasta and not just the words 'pizza' and 'pasta' themself.

Figure 6: Query restaurant - Very Italian Pizza

Figure 7: Recommended restaurant concepts model - Happy Italy

Figure 8: Recommended restaurant word embeddings model - La Bella Rosa

5.3 Image embeddings model results

The mean cosine similarity and the standard deviation excluding missing values are 0.226 and 0.095 for atmosphere, 0.149 and 0.068 for food, and 0.319 and 0.079 for food atmosphere respectively. There are again large differences between the mean values of the cosine similarity. To account for these differences, the cosine similarities are standardized again.

On average, the model improves its accuracy when standardizing the cosine similarity values (see table 10), and thus the standardized model is again preferred to the non-standardized model. The standardized image embeddings model is extremely accurate in making recommendations for cheap restaurants with an accuracy of 93%. The model recommended an expensive restaurant for only 13 of the 169 cheap restaurants. Even for the cuisine and price category combinations the model still achieves an accuracy of 71%, which is of course much higher than the expected accuracy without a model. It performs especially well for the Asian cheap and French expensive categories, but it also performs quite a bit above the expected accuracy for the Italian cheap and Asian expensive categories. The accuracy of the model for Asian restaurants is 20% better than for Italian restaurants, but this difference is almost the same for the expected accuracies.

Table 10: Accuracies of the non-standardized and standardized image embeddings model

	Category	Accuracy random sample	All concepts without standardization	All concepts with standardization
Price	Cheap	70%	91%	93%
	Expensive	30%	75%	70%
	Weighted average	*58%*	*86%*	*87%*
Cuisine	Asian	46%	78%	88%
	French	29%	69%	70%
	Italian	25%	61%	68%
	Weighted average	*36%*	*71%*	*78%*
Cuisine and price	Asian cheap	41%	75%	87%
	Asian expensive	5%	33%	33%
	French cheap	6%	0%	7%
	French expensive	23%	79%	77%
	Italian cheap	23%	63%	66%
	Italian expensive	1%	0%	0%
	Weighted average	*29%*	*65%*	*71%*

With the accuracies of this model being so high, it may be more interesting to look at the incorrect recommendations that the model makes. In figure 9 it becomes apparent that the model understands how the restaurants are structured and how the food is presented, but it lacks in recognizing what the food actually is. The trays in which the ice cream are presented is similar to the way the food of Asian cheap restaurant Kaminah Exotic is presented. It also understands the restaurants' shop-like structure. However, the model does not fully understand the difference between the Asian food and the Italian ice cream in the trays in the display case. For the French expensive Restaurant DUTCH! and Italian cheap Vapiano in figure 10, the model correctly recognizes that there is much seating area spread across the restaurant and it also recognizes the high tables. However, the sense of luxury in Restaurant DUTCH! with comfortable leather chairs, tablecloths and chandeliers is missed by the model and thus Vapiano is recommended where there are simple wooden chairs, and the styling is just clean and basic. Thus, the model seems to be missing details in pictures and focuses on similarity in the broad picture.

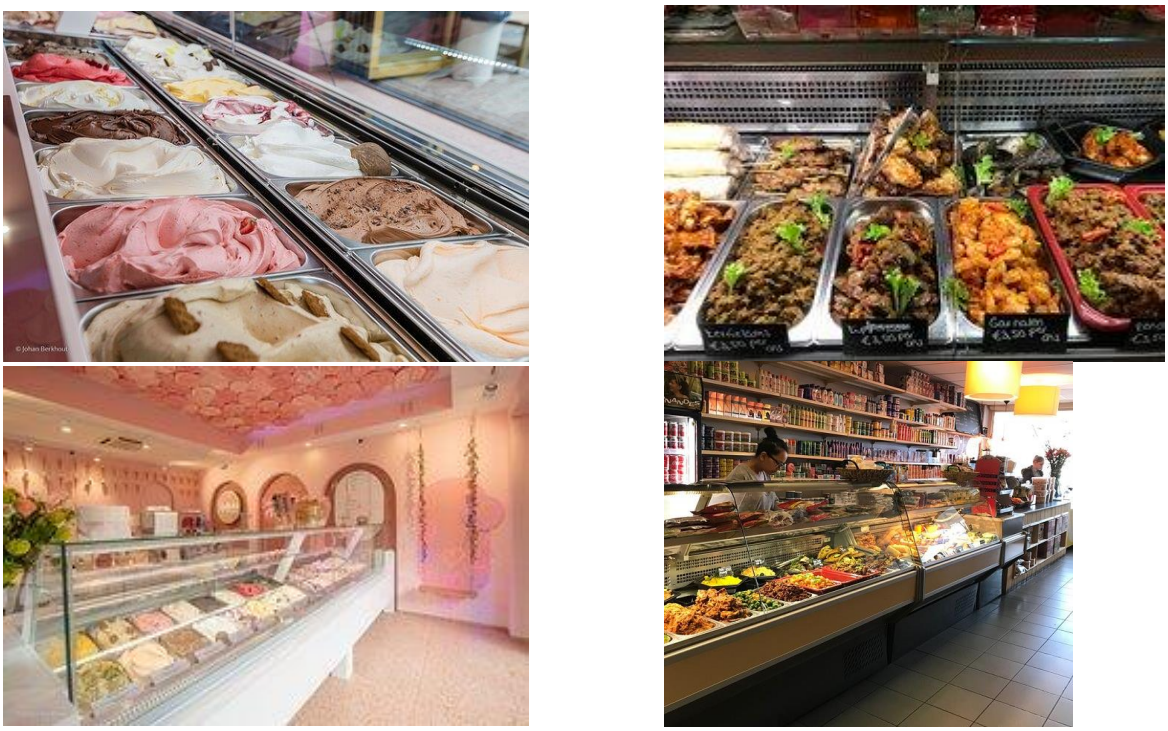

Figure 9: IJssalon IJs & Zo Wassenaar (left) with its incorrect recommendation Kaminah Exotic (right)

Figure 10: Restaurant DUTCH! (left) with its incorrect recommendation Vapiano (right)

5.4 Comparison of all models

All accuracies of the standardized versions of the three models are aggregated in table 11, to compare the different models to one another easily. What stands out is the relatively low accuracy of the word embeddings model and the relatively high accuracy of the image embeddings model. As mentioned before, the word embeddings model is less precise than the concepts model due the fact that similarity of words is measured less specifically. The image embeddings model has quite high accuracies, but the distribution of price category accuracies is much more skewed than the accuracies of the concepts model.

Even though the image embeddings model performs the best, there is still room for improvement. As seen in section 5.3, the model has difficulty in detecting the exact type of food that is served and details in broad pictures. These details seem to be important in recommending the correct price category and the type of food seems to be important in recommending the type of cuisine. Still, the model does perform better than the previous two model as Belicio Cheatday Dordrecht (figure 4, pictures of the left) had a French expensive restaurant as recommendation in both models, but the image embeddings model finally correctly recommended Vapiano (figure 10, pictures on the right), which has a much more similar atmosphere. However, the image embeddings model also made incorrect recommendations for query restaurants for which the concepts model and word embeddings model did have correct recommendations. An example of this is IJssalon IJs & Zo Wassenaar (figure 9, pictures on the left), where the first two models correctly recommended a different ice cream shop. For these models, the concept of ice cream was correctly recognized by the food recognition models, which positively facilitated the recommendation. Combining both concepts and image embeddings in one model may lead to even higher accuracies, as each model recognizes things that the other model does not. However, the concepts model's incorrect recommendation in figure 5, is also made by the image embeddings model, thus not all restaurants will receive a correct recommendation.

Table 11: Accuracies of all (standardized) models compared

	Category	Accuracy random sample	Accuracy concepts model	Accuracy word embeddings model	Accuracy image embeddings model
Price	Cheap	70%	79%	74%	93%
	Expensive	30%	79%	79%	65%
	Weighted average	*58%*	*79%*	*75%*	*87%*
Cuisine	Asian	46%	79%	71%	88%
	French	29%	70%	66%	70%
	Italian	25%	56%	44%	68%
	Weighted average	*36%*	*71%*	*63%*	*78%*
Cuisine and price	Asian cheap	41%	76%	69%	87%
	Asian expensive	5%	33%	33%	33%
	French cheap	6%	0%	0%	7%
	French expensive	23%	73%	63%	77%
	Italian cheap	23%	55%	39%	66%
	Italian expensive	1%	0%	0%	0%
	Weighted average	*29%*	*63%*	*54%*	*71%*

All models also recommended Spikkels, a French cheap restaurant, to Janssens IJssalon, an Italian cheap restaurant (see figure 11). This is viewed as an incorrect recommendation due to the different cuisines, but they are both ice cream shops. Spikkels is not just an ice cream shop as it also serves different foods such as macarons. However, people who like Janssens IJssalon will most likely also enjoy the ice cream at Spikkels, thus it is debatable how incorrect this recommendation truly is. More of these disputable recommendations may be present, which raises the question how many recommendations are classified as incorrect while people may still view it as a correct recommendation.

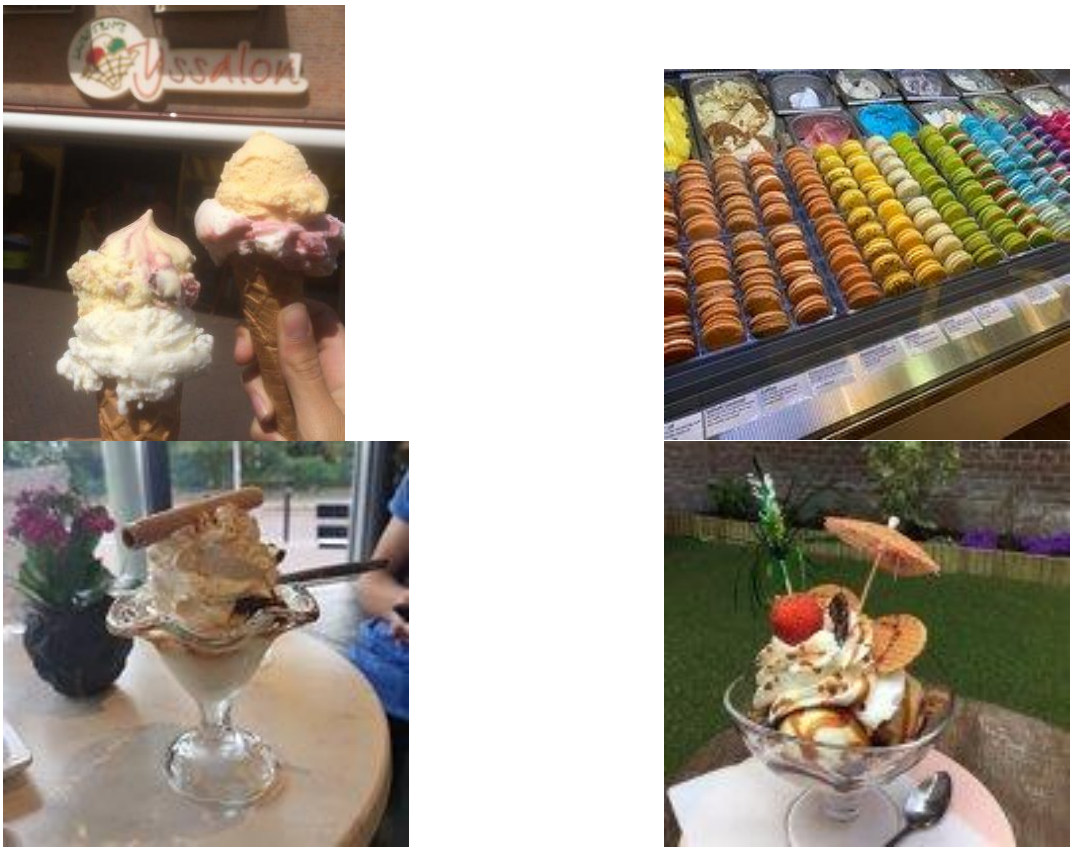

Figure 11: Janssens IJssalon (left) with its incorrect recommendation Spikkels (right)

5.5 Atmosphere versus food

In section 5.3 it seemed that details in broad atmosphere photos are important in determining the price, and the food that is recognized in food photos seemed to be important for the classification of the type of cuisine. To determine if this is correct, all models are run comparing only the atmosphere, food or food atmosphere similarity. These results are presented in table 12. For the concepts model, the food concepts are performing better on all accounts. The food concepts are relatively more important in determining the cuisine and the price compared to the atmosphere concepts with a difference in accuracy of 5% for the price category compared to a difference of 17% for the cuisine category. The food atmosphere concepts are only slightly less accurate than the food concepts. This pattern is also visible in the word embeddings model. In the image embeddings model the pattern changes and the food photos that are embedded by the general embedder instead of the food embedder now lead to a higher accuracy. The food photos that are embedded by the food embedder also lead to an extremely skewed distribution of accuracy in the price category where the cheap has an almost perfect accuracy, while the expensive restaurants have an accuracy of around 60%. The accuracy for cuisine is also very skewed with a much higher accuracy for Asian than for French and Italian restaurants. In conclusion, it seems that food does influence the cuisine accuracy more than the atmosphere, apart from the fact that food generally leads to higher accuracies.

Table 12: Accuracies of different versions of the three general models

		Concepts model			**Word embeddings model**			**Image embeddings model**		
	Category	Only atmosphere	Only food	Only food atmosphere	Only atmosphere	Only food	Only food atmosphere	Only atmosphere	Only food	Only food atmosphere
Price	Cheap	79%	83%	82%	76%	82%	81%	86%	98%	93%
	Expensive	61%	66%	69%	52%	69%	75%	56%	63%	77%
	Weighted average	*73%*	*78%*	*78%*	*69%*	*78%*	*79%*	*77%*	*88%*	*88%*
Cuisine	Asian	63%	82%	80%	56%	78%	67%	73%	87%	83%
	French	50%	64%	64%	40%	66%	64%	47%	56%	77%
	Italian	31%	47%	46%	37%	46%	36%	42%	58%	58%
	Weighted average	*51%*	*68%*	*63%*	*47%*	*66%*	*58%*	*58%*	*71%*	*75%*
Cuisine and price	Asian cheap	60%	79%	68%	46%	72%	65%	73%	91%	84%
	Asian expensive	42%	42%	42%	25%	58%	25%	33%	7%	17%
	French cheap	7%	7%	7%	7%	7%	7%	0%	0%	14%
	French expensive	50%	64%	62%	36%	68%	68%	50%	66%	86%
	Italian cheap	29%	48%	45%	32%	46%	32%	43%	61%	61%
	Italian expensive	0%	0%	0%	0%	0%	0%	0%	0%	0%
	Weighted average	*45%*	*61%*	*55%*	*37%*	*60%*	*52%*	*53%*	*68%*	*70%*

6 Conclusion and discussion

The results of the previous chapter provide evidence that images can be used to create an accurate recommendation system. The image embeddings provide the most accurate results, but these are less interpretable than the concepts model. The concept of vectors in a multi-dimensional space remains an abstract concept. The concepts of the concepts model however, are far more interpretable. Pictures of expensive restaurants contain silverware, tablecloths, more expensive ingredients, such as oysters and halibut, and have a feeling of indulgence and luxury. This is compared to pictures of cheap restaurants that can look like a diner or a pizzeria, serve fried rice and a casserole, and have more of a homemade feel. The Asian cuisine is typically recognized by people and antiques, serving rice, spring rolls and kebab, with often a shop- or market-like interior. The French cuisine serves tuna tartare and macarons on silverware with typically no people in the picture. Italian restaurants are typically either a pizzeria or an ice cream shop. Coffee is also associated with the Italian cuisine. The word embeddings model, even though it provided the least accurate results, does bring something interesting to the table, namely the broader interpretation of similarity. This model allows for recommendations that fall in the same category as the query restaurant without being nearly identical to one another. This might be beneficial for a recommendation system that also wants to provide recommendations that are just slightly different from the query product/service.

The research question '*What restaurant features are important in a restaurant recommendation system for TripAdvisor which uses image data of restaurants in the Netherlands?'* is a tough question to answer, as it depends on the type of recommendation system you want. To obtain the most accurate results the image embeddings model is the most suitable, however this model is the least interpretable. The model seems to be focussing on shapes in the picture and how the restaurant is constructed. It seems to have difficulty in recognizing the type of food if it is displayed in a similar way. Here the way the food is displayed is important in the model. Also, the layout of the restaurant plays an important role, where details such as the table setting are disregarded more. When looking at the type of pictures that are important, the food pictures with a focus on the atmosphere around the food leads to the most accurate results in the image embeddings model. Thus, the food atmosphere is the most important in this model. For the concepts model the restaurant features are often an object or an ingredient/dish, due to the concepts that can be recognized by the model. The differences between cheap and expensive restaurants and between the three different cuisines seem to lie in the examples mentioned above. In this model, the most accurate results are produced by the food concepts, thus these features of the restaurant seem to be playing the most important role in the recommendations. This is also the case for the word embeddings model, but again, this model measures the similarity between concepts more broadly. The concepts do not have to be exactly the same, they should just be

similar in meaning, allowing for slightly more diverse recommendations as restaurants do not have to be identical.

This research has a few limitations due to time constraints. The number of restaurants is quite low with only 240 restaurants and the number of scraped pictures could be higher with a maximum of 10 pictures per restaurant. More data will most likely lead to more accurate results. It was not possible to create and train an image recognition model, thus these had to be pre-trained on other data. A more restaurant-focused image recognition model might provide concepts that are more relevant for a restaurant recommendation system. The restaurant pictures are also sometimes of low quality which results in pixelated and/or dark pictures which might hinder the image recognition models. Some pictures are also not useful as the same object or dish is photographed twice. Handling missing values is often an issue in recommender systems, and here the problem is no different. There were many restaurants with less than 6 pictures and quite a few restaurants had only pictures of either the atmosphere or the food.

For future research it would be interesting to look further into a combination of text data and image data to determine recommendations for restaurants. For a focus more specifically on an image-based recommendation system, it would be interesting to look into combining multiple image recognition models to combine both the concepts and the image embeddings into one model in an attempt to use the best of both models. The condition of at least 6 pictures is, even though some experimenting has been done, quite arbitrary, so it would be interesting to further experiment with different numbers of pictures. Experimenting with different similarity measures might also be interesting and may lead to more accurate results.

References

Aggarwal, C. C. (2016). *Recommender systems* (Vol. 1). Cham: Springer International Publishing.

Al-Ghuribi, S. M., & Noah, S. A. M. (2019). Multi-criteria review-based recommender system–the state of the art. *IEEE Access, 7*, 169446-169468

Albawi, S., Bayat, O., Al-Azawi, S., & Ucan, O. N. (2017). Social touch gesture recognition using convolutional neural network. *Computational Intelligence and Neuroscience, 2018*, 1-10.

Ay, B., Aydın, G., Koyun, Z., & Demir, M. (2019). A visual similarity recommendation system using generative adversarial networks. *In 2019 international conference on deep learning and machine learning in emerging applications (Deep-ML) (IEEE)*, 44-48.

Bell, S., & Bala, K. (2015). Learning visual similarity for product design with convolutional neural networks. *ACM transactions on graphics (TOG), 34*(4), 1-10.

Bigne, E., Chatzipanagiotou, K., & Ruiz, C. (2020). Pictorial content, sequence of conflicting online reviews and consumer decision-making: The stimulus-organism-response model revisited. *Journal of Business Research, 115*, 403-416.

Burke, R. (2002). Hybrid recommender systems: Survey and experiments. *User modeling and user-adapted interaction, 12*, 331-370.

Coelho, J., Mano, D., Paula, B., Coutinho, C., Oliveira, J., Ribeiro, R., & Batista, F. (2023). Semantic similarity for mobile application recommendation under scarce user data. *Engineering Applications of Artificial Intelligence, 121*, 1-12.

Chu, W. T., & Tsai, Y. L. (2017). A hybrid recommendation system considering visual information for predicting favorite restaurants. *World Wide Web, 20*(6), 1313-1331.

Diehl, K., Zauberman, G., & Barasch, A. (2016). How taking photos increases enjoyment of experiences. *Journal of personality and social psychology, 111*(2), 119-140.

Dogru, T., & Pekin, O. (2017). What do guests value most in Airbnb accommodations? An application of the hedonic pricing approach. *Boston Hospitality Review, 5*(2), 1-13.

Dzyabura, D., & Peres, R. (2021). Visual elicitation of brand perception. *Journal of Marketing, 85*(4), 44-66.

Engel, J.F., Blackwell, R.D. & Miniard, P.W. (1995). *Consumer Behaviour, 8th edition*. Fort Worth, TX: The Dryden Press Harcourt Brace College Publishers

Han, H., & Ryu, K. (2009). The roles of the physical environment, price perception, and customer satisfaction in determining customer loyalty in the restaurant industry. *Journal of hospitality & tourism research, 33*(4), 487-510.

He, R., & McAuley, J. (2016). VBPR: Visual Bayesian Personalized Ranking from Implicit Feedback. *Proceedings of the AAAI Conference on Artificial Intelligence, 30*(1), 144-150

Hijazi, S., Kumar, R., & Rowen, C. (2015). Using convolutional neural networks for image recognition. *Cadence Design Systems Inc.: San Jose, CA, USA, 9*(1), 1-12.

Isinkaye, F. O., Folajimi, Y. O., & Ojokoh, B. A. (2015). Recommendation systems: Principles, methods and evaluation. *Egyptian informatics journal, 16*(3), 261-273.

Kotler, P. (1973). Atmospherics as a marketing tool. *Journal of retailing, 49*(4), 48-64.

Li, B., & Han, L. (2013). Distance weighted cosine similarity measure for text classification. *In Proceedings of the 14th International Conference on Intelligent Data Engineering and Automated Learning - IDEAL 2013, 8206*, (611-618).

Li, Z., Liu, F., Yang, W., Peng, S., & Zhou, J. (2022). A Survey of Convolutional Neural Networks: Analysis, Applications, and Prospects. *IEEE transactions on neural networks and learning systems, 33*(12), 6999–7019.

Liu, C. (2014). Discriminant analysis and similarity measure. *Pattern Recognition, 47*(1), 359-367.

Liu, L., Dzyabura, D., & Mizik, N. (2020). Visual listening in: Extracting brand image portrayed on social media. *Marketing Science, 39(*4), 669-686.

Lops, P., De Gemmis, M., & Semeraro, G. (2011). *Content-based recommender systems: State of the art and trends*. In Ricci, F., Rokach, L., Shapira, B., Kantor, P. (eds), Recommender Systems Handbook (73-105). Boston, MA: Springer.

McAuley, J., Targett, C., Shi, Q., & Van Den Hengel, A. (2015). Image-based recommendations on styles and substitutes. *In Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*, 43-52.

Meyers-Levy, J., & Zhu, R. (2008). *Perhaps the store made you purchase it: Toward an understanding of structural aspects of indoor shopping environment*. In M. Wedel & R. Pieters (Eds.), Visual marketing: From attention to action (193–224). New York: Psychology Press.

Mou, J., & Shin, D. (2018). Effects of social popularity and time scarcity on online consumer behaviour regarding smart healthcare products: An eye-tracking approach. *Computers in Human Behavior, 78*, 74-89.

Nguyen, H. V., & Bai, L. (2010). Cosine similarity metric learning for face verification. *In Asian conference on computer vision – ACCV 2010*, *6493*, (709-720).

Oliveira, B., & Casais, B. (2019). The importance of user-generated photos in restaurant selection. *Journal of Hospitality and Tourism Technology, 10*(1), 2-14.

O'Shea, K., & Nash, R. (2015). An introduction to convolutional neural networks. *arXiv preprint arXiv:1511.08458*, 1-11.

Pittman, M., & Reich, B. (2016). Social media and loneliness: Why an Instagram picture may be worth more than a thousand Twitter words. *Computers in Human Behavior, 62*, 155-167.

Reavey, Paula (2011), "The Return to Experience: Psychology and the Visual," in Visual Methods in Psychology: Using and Interpreting Images in Qualitative Research, Chap. 1. London: Routledge.

Rozin, P., & Royzman, E. B. (2001). Negativity bias, negativity dominance, and contagion. *Personality and social psychology review, 5*(4), 296-320.

Ruiz-Mafe, C., Chatzipanagiotou, K., & Curras-Perez, R. (2018). The role of emotions and conflicting online reviews on consumers' purchase intentions. *Journal of Business Research, 89*, 336-344.

Sarwar, B., Karypis, G., Konstan, J., & Riedl, J. (2001). Item-based collaborative filtering recommendation algorithms. *In Proceedings of the 10th international conference on World Wide Web*, 285-295.

Singhal, A. (2001). Modern information retrieval: A brief overview. *IEEE Data Eng. Bull., 24*(4), 35-43.

Sudha, M., & Sheena, K. (2017). Impact of influencers in consumer decision process: the fashion industry. *SCMS Journal of Indian Management, 14*(3), 14-30.

Sundar, S. S. (2008). *The MAIN model: A heuristic approach to understanding technology effects on credibility*. Cambridge, MA: MacArthur Foundation Digital Media and Learning Initiative.

Teubner, T., Hawlitschek, F., & Dann, D. (2017). Price determinants on Airbnb: How reputation pays off in the sharing economy. *Journal of Self-Governance and Management Economics, 5*(4), 53-80.

Wang, Y., Wang, S., Tang, J., Qi, G., Liu, H., & Li, B. (2017). CLARE: A joint approach to label classification and tag recommendation. *In Proceedings of the AAAI Conference on Artificial Intelligence, 31*(1), 210-216.

Wang, Z. J., Turko, R., Shaikh, O., Park, H., Das, N., Hohman, F., ... & Chau, D. H. P. (2020). CNN explainer: learning convolutional neural networks with interactive visualization. *IEEE Transactions on Visualization and Computer Graphics, 27*(2), 1396-1406.

Wu, X., He, R., Sun, Z., & Tan, T. (2018). A light CNN for deep face representation with noisy labels. *IEEE Transactions on Information Forensics and Security, 13*(11), 2884-2896.

Xia, P., Zhang, L., & Li, F. (2015). Learning similarity with cosine similarity ensemble. *Information sciences, 307*, 39-52.

Yu, L. C., Wang, J., Lai, K. R., & Zhang, X. (2017). Refining word embeddings for sentiment analysis. *In Proceedings of the 2017 conference on empirical methods in natural language processing,* 534-539.

Yu, W., Zhang, H., He, X., Chen, X., Xiong, L., & Qin, Z. (2018). Aesthetic-based clothing recommendation. *In Proceedings of the 2018 world wide web conference*, 649-658.

Zhang, S., Lee, D., Singh, P. V., & Srinivasan, K. (2016). How much is an image worth? An empirical analysis of property's image aesthetic quality on demand at Airbnb. *In ICIS 2016 Proceedings*, 1-20.

Zhang, M., & Luo, L. (2023). Can consumer-posted photos serve as a leading indicator of restaurant survival? Evidence from Yelp. *Management Science*, 69(1), 25-50.