FACULTY OF INFORMATION TECHNOLOGY AND ELECTRICAL ENGINEERING

Ashan Chameera Walpitage

A Food recipe recommendation system based on nutritional factors in the Finnish food community

Master's Thesis
Degree Programme in Computer Science and Engineering
June 2023

Walpitage A. C. (2023) A Food recipe recommendation system based on nutritional factors in the Finnish food community. University of Oulu, Degree Programme in Computer Science and Engineering. Master's Thesis, 152 p.

ABSTRACT

This thesis presents a comprehensive study on the relationships between user feedback, recipe content, and additional factors in the context of a recipe recommendation system. The aim was to investigate the influence of various factors on user ratings and comments related to nutritional variables, while also exploring the potential for personalized recipe suggestions. Statistical analysis, clustering techniques, and sentiment analysis were employed to analyze a dataset of food recipes and user feedback. We determined that user feedback is a complex phenomenon influenced by subjective factors beyond recipe content alone. Cluster analysis identified four distinct clusters within the dataset, highlighting variations in nutritional values and sentiment among recipes. However, due to an imbalanced distribution within the clusters, these relationships were not considered in the recommendation system. To address the absence of user-related data, a content-based filtering approach was implemented, utilizing nutritional factors and a health factor calculation. The system provides personalized recipe recommendations based on nutritional similarity and health considerations. A maximum limit of 20 recommended recipes was set, allowing users to specify the desired number of recommendations. The accompanying API also provides a mean squared error metric to assess recommendation quality. This research contributes to a better understanding of user preferences, recipe content, and the challenges in developing effective recommendation systems for food recipes.

Keywords: User feedback, recipe recommendation system, content-based filtering, clustering analysis, sentiment analysis, nutritional variables.

TABLE OF CONTENTS

FOREWORD

I am grateful to present this thesis as part of my Master's degree in Computer Science and Engineering with a specialization in Applied Computing at the University of Oulu, Finland. The focus of my research lies in the fascinating field of hybrid recommendation systems, specifically exploring the utilization of multiple information sources derived from food recipes within the Finnish food community. Furthermore, I incorporated health factors based on WHO guidelines to enhance the personalized recommendations provided by the system.

The field of recommendation systems encompasses various disciplines, including classification models, data analysis, and domain-specific expertise. The opportunity to delve into these areas and contribute to the advancement of personalized services for customers has been truly captivating.

I would like to express my deepest gratitude to my supervisor, Assistant Prof. Mourad Oussalah, for his invaluable guidance, insightful suggestions, and unwavering support throughout the thesis process. Prior to this research, I had the privilege of working with him on social media analysis using Twitter data and delving into the intricacies of Natural Language Processing. I am immensely grateful for the continuous mentorship and encouragement he provided.

I would also like to extend my heartfelt thanks to my family members for their unwavering support and encouragement throughout this journey.

It is my hope that this thesis contributes to the ever-evolving field of recommendation systems, paving the way for enhanced personalized recommendations in the domain of Finnish food recipes while considering vital health factors.

16th of June, Oulu
Ashan C. Walpitage

ABBREVIATIONS

WHO	World Health Organization
FSA	Food Standard Agency
DASH	Dietary Approaches to Stop Hypertension
NLP	Natural Language Processing
CNN	Convolutional Neural Networks
SVM	Support Vector Machines
LSTM	Long Short-Term Memory
BERT	Bidirectional Encoder Representations from Transformers
NCD	Noncommunicable Diseases
SNA	Social Network Analysis

1 INTRODUCTION

What are recommendation systems? Recommender systems play a crucial role in today's digital landscape, providing users with personalized suggestions and supporting decision-making processes. A recommender system is a subclass of an information filtering system that aims to predict user preferences for products or items [1]. These systems are designed to help users discover relevant and desirable items from a vast array of options. Two main tasks in recommender systems are ranking and rating prediction tasks [2]. The rating task is to predict the rating that a user would give to a certain product, while the ranking task aims to recommend the most relevant products to the users based on the predicted ranking ratings.

The personalization aspect of recommender systems is essential. Personalized recommender systems tailor their suggestions to each user's needs, interests, and preferences. By analyzing user behavior, preferences, and feedback, these systems can provide meaningful and relevant recommendations that resonate with users on a personal level as well as their similarity with other items. Personalization improves the user experience by making it more engaging and intuitive, which leads to greater customer satisfaction and loyalty.

Personalized recommendation systems have been essential to the success of applications in a variety of fields, including e-commerce, content distribution, food and beverages, and trip booking [3]. The huge diversity of products and services provided by these applications can be overwhelming to consumers, making it difficult to locate acceptable solutions. Personalized and content-based recommendations lessen this load by compiling a list of items or services that meet the requirements and interests of the consumers. These technologies assist users in making educated selections and navigating the overwhelming amount of options by providing the most accurate suggestions.

Personalization is an important component in improving program usability and engagement, particularly in the context of content-based recommendation systems [4]. By using user behavior and preferences to propose new or relevant content, objects, or services, these systems offer an interactive and dynamic user experience. material-based recommendation systems provide personalized recommendations that coincide with users' interests by assessing the features of the material itself and comparing it with users' preferences [5]. By giving targeted ideas that respond to customers' individual requirements, this proactive strategy not only promotes user engagement but also encourages recurrent usage. Content-based recommendation systems contribute to consumer pleasure, loyalty, and, ultimately, the long-term success of applications and platforms by giving individualized recommendations.

Why are health factors important for food recipe recommender systems? Health aspects are important in food recipe recommender systems, especially when it comes to encouraging healthy eating habits and meeting individual dietary demands. By including health criteria in these systems, users are guaranteed to obtain suggestions that are tailored to their individual health objectives, preferences, and constraints. These systems can help users' general well-being and support their attempts to

maintain a balanced and healthy diet by taking the healthiness of recommended meals into account.
Recommendation systems can use many forms of information to propose the healthiest eating selections. Some significant information sources that can be used are:

- Nutritional Information: Analyzing the nutritional content of food items is essential in determining their healthiness. This involves taking into account macronutrients (carbohydrates, proteins, and fats), micronutrients (vitamins and minerals), and overall calorie content. Recommendation systems can offer healthier alternatives based on individual dietary requirements and objectives by using nutritional information.
- Dietary Preferences and Restrictions: By taking into consideration individual dietary preferences and Restrictions, recommendation systems may provide personalized and content-based recommendations that are tailored to their specific needs. This might include catering to different diets, such as vegetarian, vegan, gluten-free, or dairy-free, as well as avoiding items that users are allergic to or wish to avoid due to health concerns.
- Health Goals and Profiles: It is critical to consider consumers' health goals and profiles while proposing the healthiest meal. Some people may want to reduce weight, manage chronic diseases such as diabetes or hypertension, or adhere to strict dietary requirements. Recommendation systems can deliver suggestions that support users' intended results by understanding their health aims from the content of the health facts.
- Ingredient Quality and source: Evaluating the quality and source of ingredients might help to make recommended recipes healthier. Recipes with fresh, organic, or locally sourced ingredients might be prioritized by recommendation algorithms, encouraging a more wholesome and healthful dining experience.
- User Ratings and Feedback: Integrating user ratings and feedback into recommendation systems enables continuous development and fine-tuning of proposed recipes. Users' ratings of recipes based on flavor, healthiness, and overall pleasure might help the algorithm promote recipes that have gotten favorable feedback and match the tastes of the users.

Personalized recommendation systems may present consumers with healthier food selections that adapt to their unique dietary needs, tastes, and health objectives by taking into account these diverse information sources. These systems help to promote a healthy, balanced diet, which improves consumers' overall well-being.

What kind of information can be used in recommending the healthiest food?
Personalized content-based recommendation systems can use numerous types of information to make important ideas when proposing the healthiest eating alternatives. These systems can give suggestions that match users' health objectives and promote healthy eating habits by taking into account several criteria relating to nutritional content, dietary requirements, and ingredient quality.

- Nutritional Composition: Content-based recommendation systems can use nutritional composition information to propose the healthiest solutions. This includes taking into account macronutrients (carbohydrates, proteins, fats) and micronutrients (vitamins, minerals) in the recipes. These systems can recommend food selections that match users' individual dietary demands and contribute to a balanced and nutritious diet by assessing nutritional information.
- Dietary Guidelines: Incorporating dietary guidelines into recommendation systems guarantees that proposed food selections adhere to recognized health and nutrition criteria. These guidelines may include general health organization advice or specialized dietary programs such as the Mediterranean diet or DASH (Dietary Approaches to Stop Hypertension). Content-based recommendation systems can promote better eating choices by following these recommendations.
- Allergies and dietary limitations: It is critical to consider users' allergies and dietary limitations when offering the best meal alternatives customized to individual requirements. Users may be allergic to certain substances or adhere to certain dietary habits, such as gluten-free, dairy-free, or vegetarian/vegan diets. By taking these constraints into consideration, recommendation systems may make suggestions that avoid allergies or adhere to specific dietary preferences, assuring the safety and well-being of users.
- Nutritional Value of Individual Components: Analyzing the nutritional value of individual ingredients is important in advising better diet selections. Recipes with nutrient-dense components, such as fruits, vegetables, whole grains, lean meats, and healthy fats, might be prioritized by recommendation algorithms. Users can obtain ideas that help to a healthful and balanced diet by promoting dishes that incorporate these beneficial items.
- Cooking Methods and Techniques: Cooking methods and techniques that enhance healthy food preparation might be considered by recommendation systems. Recipes that use steaming, grilling, roasting, or sautéing instead of deep-frying or utilizing excessive quantities of oil may be suggested. These systems can help users' general well-being by offering recipes that highlight better cooking procedures **[6]**.

1.1 Research Questions

Research Question 1: A roadmap for comprehending the occurrence of correlation in food datasets.

a. What is the correlation between different nutritional factors of the foods and their popularity in social media?

Research Question 2: A statistical analysis showing the dominant features that influence user feedback in the food recommendation dataset

a. How can the influence of different features on user feedback be measured and analyzed statistically?
b. Are there specific features that have a stronger impact on user feedback compared to others?
c. What are the potential implications of the dominant features on the accuracy and effectiveness of food recommendation systems?

Research Question 3: How can a hybrid food recommender system be designed to incorporate food content, nutrition content, and health factor?

a. What are the existing approaches for calculating the health factor of the foods?
b. How can the health factor of recommended foods be taken into account in the hybrid system?
c. What methods can be used to effectively capture and incorporate user preferences in the hybrid system?

1.2 Contribution

- Generate an attributed graph based on the nutritional properties of food recipes:

  In order to capture the intricate relationships between food recipes and their nutritional properties, we developed an approach to generate an attributed graph. This graph represents each recipe as a node, with edges connecting related recipes based on their nutritional properties and other properties like preparation time. By considering attributes such as energy, protein, carbohydrates, fat, saturated fat, dietary fiber, salt, preparation time, and difficulty level, we were able to create a comprehensive representation of the recipe dataset, enabling more accurate analysis and recommendation generation.

- Data pre-processing step for calculating the sentiment of the foods:

  Pre-processed the incorrect JSON types and add missing values parameters. Also remove the incomplete data for the dataset. To ensure that our analysis encompassed a wider audience and to facilitate a more comprehensive understanding of user feedback, we employed a data pre-processing step to translate all user comments from Finnish to English. By leveraging the power of Google Translate, we were able to overcome the language barrier and include a diverse range of user opinions in our analysis using VADER. This step allowed us to gain sentiments toward various food recipes, enhancing the overall effectiveness using VADER.

- Extract sentiment scores of the food recipes using Afinn and Vader:

  To gain a deeper understanding of the sentiments associated with the food recipes in our dataset, we utilized two popular sentiment analysis tools: Afinn and Vader. These tools provided us with sentiment scores that classified each recipe's sentiment as positive, neutral, or negative. By extracting sentiment scores, we were able to uncover valuable insights into the overall sentiment trends surrounding different recipes. This information played a crucial role in refining our recommendation algorithm and classification biases regarding recipe properties.

- Clustering the foods in our dataset to find different clusters and their properties using k-means and spectral clustering:

  To discover inherent patterns and structures within the recipe dataset, we employed clustering techniques such as k-means and spectral clustering. These algorithms allowed us to group similar recipes together based on their shared properties, such as energy, protein, carbohydrates, fat, saturated fat, and other nutritional content. By identifying distinct clusters, we gained a better grasp understanding of the diversity within the dataset and the unique characteristics of each cluster. This knowledge served as a foundation for further analysis and recommendation generation.

- Generate cosine similarity matrices and normalize them:

  By calculating cosine similarity matrices, we quantified the similarity between pairs of food recipes based on their various attributes. These matrices provided a measure of similarity, allowing us to identify recipes with comparable nutritional profiles or other shared characteristics. To visualize these relationships, we utilized Gephi, a powerful graph visualization tool. Through this process, we generated a graph representation that showcased the interconnectedness of recipes, with edges representing similarities. This visualization facilitated a more intuitive understanding of the relationships between different recipes.

- Find correlation and coefficient values in different characteristics:

  To uncover potential biases and correlations within the dataset, we conducted an in-depth analysis of the relationships between different properties of food recipes. By calculating correlation coefficients, we identified associations between variables such as energy level, macronutrient ratios, and user preferences. This analysis helped us better understand the underlying factors that influence recipe selection and satisfaction. By accounting for these biases and correlations, we aimed to develop a recommendation system that caters to a balanced and nutritious food selection.

- Hybrid recommender system based on food properties and Health factors:

Building upon the insights gained from our analysis, we developed a hybrid recommender system that incorporated both food properties and factors related to the Food Standard Agency (FSA). By combining nutritional attributes, and FSA

guidelines, our recommendation system provided users with healthy recipe suggestions. This unique approach ensured that the recommendations not only aligned with individual preferences but also adhered to established health guidelines. By integrating these factors, we aimed to promote balanced nutrition and support users in making informed and wholesome food choices.

1.3 Structure of Thesis

The structure of the thesis is outlined as follows. In Section 2, an extensive review of relevant literature is presented, focusing on clustering, sentiment analysis, classification, and recommendation systems. This section provides a comprehensive understanding of the existing research in these areas. Section 3 delves into the methods, concepts, and datasets utilized in the study, providing insights into the practical aspects of the research. Section 4 showcases the results obtained from the analysis and facilitates in-depth discussions, offering critical interpretations and implications. Section 5 concludes the thesis by providing a summary of the key findings, emphasizing their significance and potential contributions to the field. Additionally, the thesis incorporates an appendix, specifically Section 7, which contains a URL link to a public GitHub repository. This repository houses the code developed as part of this research, enabling easy access and future exploration of the implemented algorithms and techniques. The well-structured organization of the thesis facilitates a coherent and systematic presentation of the research, enabling readers to navigate through the different sections seamlessly and gain a comprehensive understanding of the study.

2 RELATED WORK

2.1 Recommendation Systems

A recommendation system is a sort of data filtering system that predicts or proposes goods, products, or content that consumers may be interested in. To provide tailored suggestions, it examines user preferences, activity, and other pertinent data. They are classified into three types: collaborative filtering, content-based filtering, and hybrid methods, each of which uses distinct algorithms and strategies to produce suggestions.

2.1.1 Content-Based Filtering

Content-based filtering is a prominent approach in recommendation systems for providing customized recommendations based on the properties and attributes of entities as well as the user's preferences. In contrast to collaborative filtering approaches that rely on user behavior and user similarities, content-based filtering focuses on the inherent qualities and substance of the objects themselves.

The basic concept underlying content-based filtering is to assess item content and attributes in order to create item profiles or representations. Depending on the domain of the recommendation system, these item profiles collect crucial qualities such as textual descriptions, metadata, genre, actors, directors, and other relevant attributes. The system can propose entities that are similar in content to those the user has enjoyed in the past by comparing the item profiles with the user's preferences or profile.

The recommendation system generally takes a number of stages to accomplish content-based filtering. First, it collects information about the products, such as textual descriptions, metadata, user-generated tags, ratings, and other pertinent information. This data is then pre-processed in order to extract useful characteristics and represent the objects in an analysis-ready way.

One of the common techniques used in content-based filtering is the creation of item profiles through feature extraction. This entails extracting relevant information such as keywords, genres, or other properties from the item data. In a movie recommendation system, for example, the features may include the movie genre, director, actors, and storyline keywords. These characteristics are used to build a profile or representation of each object.

Following the creation of the item profiles, the system produces a user profile based on the user's preferences or prior interactions. The user profile often includes information on the user's chosen traits, such as genres, actors, or other relevant attributes. This user profile is used to compare and match with the item profiles.

The system employs a similarity or relevance metric to propose items to the user. This metric compares the user profile to the item profiles to assess the degree of similarity between the user's preferences and the item content. Depending on the kind of features and data format, several similarity metrics, such as cosine similarity, Jaccard similarity, or Euclidean distance, can be used.

The recommendation algorithm evaluates the products based on their closeness to the user profile, and the items with the highest rankings are suggested to the user. Additional elements, such as the user's verbal input or implicit signals generated from

their behavior, can be used to modify the amount of personalization. For example, to enhance recommendations, the algorithm may assign greater weight to the user's favored genres or use the user's ratings of related entities.

In recommendation systems, content-based filtering has various advantages. Because the emphasis is on the item's substance rather than on past user statistics, it may make suggestions even for new or unpopular products. Because the suggestions are based on specific item qualities and attributes, it also provides straightforward explanations of recommendations. Furthermore, content-based filtering can help with the cold start problem, which occurs when new users have little or no user data.

However, there are several limits to content-based screening. It is strongly reliant on the accessibility and quality of item content and qualities. The suggestions may not be ideal if the item data is insufficient, erroneous, or lacking key attributes. Information-based filtering is also susceptible to the "filter bubble" effect, in which users may obtain suggestions that reinforce their existing preferences while limiting their exposure to different information.

Various strategies and developments have been presented to improve the efficiency of content-based filtering. Textual data may be analyzed using Natural Language Processing (NLP) techniques to extract semantic meaning from item descriptions or reviews. Sentiment analysis may be used to take into account user sentiment toward products and deliver more sophisticated suggestions. Decision trees, support vector machines, and neural networks are examples of machine learning techniques that may be used to understand complicated patterns and increase recommendation accuracy.

In summary, content-based filtering is an important approach in recommendation systems that uses the content and qualities of entities to generate tailored suggestions. Content-based filtering allows the system to propose products that fit with the user's interests by assessing intrinsic qualities and comparing them to the user's preferences. While it has limits, advances in data processing, machine learning, and natural language processing (NLP) continue to improve the efficacy and usefulness of content-based filtering in a variety of disciplines [7].

2.1.2 Collaborative Filtering

Collaborative filtering is a popular approach in recommendation systems that makes individualized recommendations by relying on the collective intellect and actions of users. To create suggestions, collaborative filtering investigates the relationships and interactions between people and entities, as opposed to content-based filtering approaches that focus on item features.

The basic idea underlying collaborative filtering is that users who have shared preferences in the past are likely to share them in the future. Collaborative filtering seeks to propose entities to a user based on the preferences and behavior of other like-minded users by utilizing this similarity.

The recommendation system generally takes a set of stages to achieve collaborative filtering. The first phase entails gathering and aggregating information about user-item interactions, such as ratings, reviews, purchase history, or click-through data. This information gives insights into user preferences and habits, which are critical for making reliable suggestions. There are two sorts of collaborative filtering approaches: memory-based and model-based methods.

Memory-based collaborative filtering algorithms, also known as neighborhood-based methods, find related people or objects based on the acquired user-item data. These approaches employ similarity measurements, such as cosine similarity or Pearson correlation coefficient, to determine how similar users or products are. The similarity metric indicates how similar two people or goods' preferences are. Following the calculation of similarity between users or objects, the system discovers a neighborhood of similar users or items and uses their preferences to provide suggestions. If User A and User B have similar tastes for multiple products, the system may suggest items liked by User B to User A.

Model-based collaborative filtering approaches, on the other hand, develop a model or algorithm from the user-item data to capture the underlying patterns and correlations in the data. These models may be built using a variety of machine-learning approaches, including matrix factorization, Bayesian networks, and neural networks. Based on the learned patterns, the model is trained on user-item data and learns to forecast user preferences or produce suggestions. A matrix factorization approach, for example, can factorize the user-item interaction matrix into latent components indicating user and item preferences, allowing individualized suggestions to be generated.

Both memory-based and model-based collaborative filtering systems offer benefits and drawbacks. Because they rely on direct comparisons between users or entities, memory-based approaches are straightforward to implement and interpret. They can also successfully manage new users or products because they simply require similarity metrics. However, when the size of the user-item matrix expands, these techniques may have scalability concerns. By identifying complex patterns in the data, model-based approaches can manage big datasets and deliver reliable suggestions. However, they may demand more computing resources and experience to efficiently construct and train the models.

In recommendation systems, collaborative filtering has various advantages. Because it does not rely on explicit item content or attributes, it is applicable to a wide range of domains and item kinds. It may propose both popular and obscure entities by leveraging user collective intelligence. By depending on the preferences of comparable users or products, collaborative filtering overcomes the cold start problem, which occurs when there is little or no knowledge about new people or items.

However, collaborative filtering systems have several disadvantages. They require a large volume of user-item interaction data to create appropriate suggestions. When data is sparse or there is a lack of user feedback, collaborative filtering may struggle to produce relevant recommendations. Collaborative filtering approaches may also suffer from "popularity bias," in which popular entities receive more suggestions, overshadowing potentially relevant but less-known items. Furthermore, collaborative filtering might be vulnerable to shilling attacks or the manipulation of user profiles to affect suggestions.

Various strategies and advances have been presented to improve the efficacy of collaborative filtering. To capitalize on the benefits of diverse methods, hybrid systems integrate collaborative filtering with other recommendation techniques such as content-based filtering or demographic information. Trust-based collaborative filtering uses user trust or social ties to impact suggestions. Context-aware collaborative filtering considers contextual elements like time, location, or device to make more appropriate choices.

The application of machine learning and deep learning algorithms has enhanced collaborative filtering in recent years. Deep neural networks, recurrent neural networks (RNNs), and convolutional neural networks (CNNs) have demonstrated promise in collecting complicated patterns in user-item data and producing more accurate suggestions.

Finally, collaborative filtering is a strong method in recommendation systems that harnesses users' collective expertise and actions to deliver individualized suggestions. Collaborative filtering finds comparable people or objects and uses their preferences to provide suggestions by evaluating user-item interactions. Despite its limits, collaborative filtering is evolving in tandem with advances in machine learning and data processing, allowing for more accurate and effective recommendation systems [8] [9].

2.1.3 Hybrid-Based Filtering

To overcome the limits of individual approaches and give more accurate and tailored suggestions, hybrid-based filtering algorithms combine the characteristics of numerous recommendation systems. Hybrid techniques try to harness the complementing characteristics of each methodology and increase overall suggestion quality by merging diverse filtering methods.

The impetus for hybrid filtering stems from the observation that no one suggested approach is optimal in all cases. Different approaches each have their own set of advantages and disadvantages, and by combining them, it is possible to improve recommendation performance, solve data sparsity difficulties, deal with the cold start problem, and deliver more diversified and tailored suggestions.

There are numerous techniques to developing hybrid recommendation systems, and the strategy used relies on the features of the data, available recommendation algorithms, and intended recommendation goals. In this section, we will look at several typical hybrid strategies used in recommendation systems:

- Weighted Hybridization:

This strategy combines recommendations in several ways by allocating weights to each method depending on its performance or relevance. Weights can be modified statically or dynamically based on circumstances or user preferences. If collaborative and content-based filtering methods are utilized, for example, the suggestions from each approach can be weighted and combined to generate the final recommendation list.

- Feature Combination:

This strategy entails combining characteristics or attributes from several recommendation systems into a single model. In a hybrid system that combines collaborative filtering with content-based filtering, for example, the user and item features from each technique can be integrated into a single feature representation. The combined representation may then be used by machine learning algorithms to

discover the appropriate weights for individual attributes and create recommendations.

- Switching Hybridization:

Different recommendation algorithms are used for different contexts or user segments in switching hybridization. Depending on the conditions or criteria, the system alternates between techniques. Collaborative filtering, for example, maybe more successful for individuals with a rich engagement history, whereas content-based filtering may be better suited for new users with minimal data. Based on the user's profile or behavior, the system chooses the best approach to utilize.

- Cascade Hybridization:

Cascade hybridization includes pre-filtering the item space with one recommendation approach and then refining the suggestions with another method. For example, as the first phase, a content-based filtering mechanism can be used to choose a subset of entities related to the user's preferences. The improved item set may then be utilized to produce more customized suggestions via collaborative filtering.

- Ensemble Methods:

In hybrid filtering, ensemble methods aggregate the outputs of numerous recommendation algorithms using techniques such as voting, averaging, or stacking. Each technique provides its own set of suggestions, which are then aggregated by the ensemble algorithm to make the final recommendation list. The goal of ensemble techniques is to harness the knowledge of numerous approaches while improving the resilience and accuracy of suggestions.

- Meta-level Hybridization:

In meta-level hybridization, the suggestions from several approaches are fed into a higher-level model, which mixes and refines the outputs. A machine learning algorithm, a rule-based system, or even a human expert can be used to create this meta-level model. Based on historical data or user feedback, the meta-model learns to weight or combine recommendations from multiple approaches, resulting in an optimum recommendation.

Filtering approaches based on hybrids have various benefits over standalone methods. By combining numerous sources of information, they can overcome the restrictions of data sparsity and the cold start problem. Because multiple methods capture different parts of user preferences and item attributes, hybrid systems can deliver more diversified suggestions. Furthermore, hybridization enables flexibility and adaptability since the system may dynamically modify the combination of approaches in response to user feedback or changing requirements.

On the other hand, designing and implementing hybrid recommendation systems might be difficult. It necessitates careful consideration of the integration process, the selection of recommendation techniques, and performance evaluation. The quality and diversity of the recommendation techniques used in hybrid systems

substantially influence their success. It is also critical to ensure correct integration and eliminate redundancy or contradictory advice.

Finally, hybrid-based filtering approaches in recommendation systems offer a strong option for improving suggestion quality, customization, and user happiness. Hybrid techniques provide more accurate, diversified, and context-aware suggestions by combining the capabilities of different methodologies. As recommendation systems advance, hybridization is anticipated to play an important role in advancing the field and addressing the improving expectations of users across several domains [10] [5].

2.1.4 Challenges and Limitations in Recommendation Systems

Recommendation systems have become an essential component of many online platforms and services, allowing consumers to enjoy personalized and personalized experiences. These systems, however, are not without their difficulties and constraints. In this section, we will look at some of the major issues that recommendation systems encounter, as well as the restrictions that might develop during their implementation [11].

- **Scalability**: As online platforms' user bases and item catalogs expand, scalability becomes a critical difficulty for recommendation algorithms. It might be computationally hard to process and analyze massive amounts of data in real-time in order to create customized suggestions. It is a difficult effort to ensure that the system can manage increased user traffic and growing datasets while remaining efficient.

- **Sparsity of data**: Recommendation systems rely on user data to provide reliable suggestions. However, user-item interaction data might be sparse, which means that the user-item matrix frequently has many missing elements. This sparsity is a problem since it restricts the system's capacity to adequately represent user preferences and item features, perhaps resulting in erroneous or less diversified suggestions.

- **Cold Start Problem**: One of the fundamental issues in recommendation systems is the cold start problem, which happens when there is inadequate data on users or goods. In such circumstances, it is difficult to make reliable suggestions due to a lack of information to comprehend user preferences or item qualities. This is especially true for new users or new entities with insufficient previous data.

- **Diversity and serendipity**: Recommendation systems frequently confront the difficulty of combining customized suggestions with diversity promotion. While the goal of customization is to give appropriate suggestions based on a user's preferences, it can accidentally lead to filter bubbles, in which

consumers are only exposed to familiar products or materials. To improve user experience, ensure serendipity, and provide various recommendations that go beyond users' established preferences.

- **Overfitting and Overspecialization**: Overfitting occurs when models become too particular to the given training data in recommendation systems. Overfitting can result in a lack of generality, making it difficult to propose entities to consumers with varying interests or preferences. Similarly, overspecialization happens when the system concentrates solely on a small number of popular products, ignoring the long tail of less popular goods that may be of interest to certain users.

- **Privacy and trust**: To deliver individualized suggestions, recommendation algorithms frequently gather and analyze user data. However, privacy problems might occur when consumers are hesitant to share personal information or are uninformed of how their data is being utilized. To address these issues and preserve user confidence in the system, it is critical to build trust with users and provide clear data management methods.

- **Evaluation and Feedback**: It is difficult to assess the performance and effectiveness of recommendation systems. Traditional assessment measures, such as accuracy or precision, may not accurately reflect overall user happiness or the system's capacity to fulfill a wide range of user demands. Gathering and implementing user feedback into the assessment process is critical for understanding the system's strengths and limits and making continual changes.

- **Dynamic and Evolving Preferences**: User preferences and item qualities can change over time. To deliver up-to-date and meaningful recommendations, recommendation systems must adapt and capture these dynamic changes. Incorporating techniques to deal with idea drift, user input, and contextual changes might assist in addressing the difficulty of changing preferences.

- **Considerations for Ethical and Bias**: Recommendation systems have the ability to affect user behavior and change their opinions. As a result, addressing ethical considerations such as algorithmic bias, fairness, and transparency is critical. Bias may be caused by a variety of factors, including biased training data, and it can result in discriminating or distorted suggestions. Building trustworthy and inclusive systems requires ensuring fairness, openness, and bias reduction in the recommendation process.

- **Cold Start Problem for Items**: Similar to the cold start problem for users, recommendation algorithms confront difficulties when dealing with new goods that have little or no historical data. It is difficult to provide reliable suggestions for these things since there is inadequate information to evaluate

their qualities or user preferences. To enable successful suggestion creation, techniques for dealing with the cold start problem for items must be developed.

In summary, although recommendation systems provide considerable benefits in terms of customizing user experiences and enabling content discovery, they also present a number of obstacles and limits. Overcoming these obstacles necessitates ongoing research and innovation, as well as a multidisciplinary strategy that tackles technological, user-centric, ethical, and scalability concerns. Understanding and overcoming these problems allows recommendation systems to improve to deliver more accurate, diversified, and trustworthy suggestions, hence increasing user happiness and engagement.

2.2 Healthy Recommendations and food recommender system

A healthy food recommender system is intended to deliver individualized suggestions for nutritious food options based on personal preferences, dietary objectives, and health concerns. A system like this one attempts to help people make more educated food choices and encourage a healthier lifestyle. A healthy food recommender system can deliver personalized suggestions that correspond with users' individual dietary needs and wellness goals by utilizing data analysis, machine learning techniques, and nutritional expertise.

A healthy food recommender system is built around a massive quantity of data, which includes nutritional information about various food products, user profiles, dietary standards, and expert knowledge on healthy eating. This data is used to construct clever algorithms that can generate relevant suggestions. Several parameters are considered by the algorithm to guarantee that the recommendations are relevant, healthy, and consistent with the user's interests.

The user's dietary objectives and limits are one of the most important factors in a healthy food recommender system. Weight management, allergies, specialized diets (e.g., vegetarian, vegan, paleo), or health issues (e.g., diabetes, hypertension) are examples of such variables. Understanding these limits allows the system to filter out meal selections that may not be appropriate and prioritize recommendations that are compatible with the user's dietary needs.

The technology examines the nutritional makeup of various food products to deliver precise and individualized recommendations. This involves looking at the macronutrients (carbohydrates, proteins, and fats) and micronutrients (vitamins and minerals) in each food. The system may analyze the nutritional content of items and their compatibility for individual dietary demands by utilizing existing nutritional databases and recommendations. The recommender system can offer items that satisfy the user's intended nutrient intake or alternatives to balance their nutritional intake based on the nutritional profiles.

Taking into consideration the user's preferences, tastes, and dietary selections is another critical part of a healthy food recommender system. This is accomplished by gathering data on the user's chosen cuisines, ingredients, flavor profiles, and cooking methods. Understanding these preferences allows the system to create

suggestions that not only meet the user's nutritional needs but also cater to their own taste preferences. This tailored approach raises the chances of consumers accepting the recommended items and establishing healthier eating habits in the long term.

A healthy food recommender system can benefit from feedback and user interactions to improve its performance. The system may continually learn and enhance its suggestions by collecting users' comments on recommended foods, such as ratings, reviews, or consumption habits. This feedback loop enables the system to adapt to changing user preferences, increase the grasp of its dietary objectives, and enhance the accuracy of future recommendations.

A complete healthy food recommender system might also include information such as food quality, sourcing procedures, and sustainability issues. This broadens the scope of the study beyond human health to include larger aspects such as environmental impact and ethical concerns about food choices. By combining these aspects, the recommender system may direct consumers toward environmentally and socially responsible food selections, so encouraging not just personal health but also environmental and social well-being.

In conclusion, a healthy food recommender system integrates nutritional knowledge, user preferences, and dietary objectives to deliver individualized suggestions for nutritious food options. The system intends to enable users to make educated food decisions, adopt healthy eating habits, and enhance their overall well-being by utilizing data analysis, machine learning algorithms, and user feedback. Healthy food recommender systems have the potential to play a key role in encouraging better lives and solving public health concerns connected to nutrition as technology progresses and more data becomes accessible [12] [13] [14].

3 IMPLEMENTATION

3.1 Dataset

The dataset used in this study was obtained from the website https://www.valio.fi/, a highly visited and prominent social media platform focused on food, with over 25 million annual visits. The dataset comprises 5,472 recipes that were posted on the website between 2012 and 2022. Each recipe in the dataset includes various attributes such as the recipe name, published time, ingredients, preparation time, difficulty level, tags, users' ratings, users' comments, and the nutritional content per 100 grams, including energy, protein, carbohydrates, fat, saturated fat, dietary fiber, and salt. In the preprocessing phase, 663 recipes that lacked nutritional information were excluded, resulting in a final dataset of 4,833 recipes. This dataset provides a comprehensive collection of recipes from the Valio Oy commercial website, along with user ratings, comments, and detailed nutritional values. It offers a rich and diverse source of information for evaluating and analyzing the relationships between recipe properties and nutritional values, making it suitable for studying hybrid recommendation systems in the food domain.

Table 1- Basic statistics of the crawled dataset

Entity	Description
Number of all recipes	5,472
Years of published recipes	2012-2022
Recipes containing complete information	4833
Number of recipes containing ratings	3197
Number of recipes containing comments	3060

3.2 Methodology

In the preliminary stage, we embarked on data preprocessing procedures to ensure the optimal handling of the dataset. The dataset consisted of user comments encapsulated within a JSON structure, encompassing a plethora of properties pertaining to the comments themselves. These properties encompassed crucial information such as the timestamp of the comments, unique identifiers, the type of user responsible for the comment's creation, the level of anonymity associated with the users, and other relevant attributes.

Our primary objective during this phase revolved around rectifying any incomplete JSON structures encountered within the dataset. These structures presented instances where the user comments were not properly structured, requiring careful restoration to align with the standardized JSON format. This meticulous repair process involved addressing missing elements, reconstructing malformed sections, and ensuring the syntactical integrity of the JSON representation.

Furthermore, a key aspect of our data preprocessing entailed the translation of the comments from their original Finnish language to English. Using the google

translator API we performed the translation. Through language translation techniques, we facilitated a seamless transition, effectively bridging the language barrier that may have impeded the analysis and interpretation of the user comments. Following diagram shows pipeline of our methodology.

Graph 1: Steps of Data Pipeline

In order to identify the key nutrition factors that exhibit significant variance and will be used for subsequent clustering techniques, a Principal Component Analysis (PCA) was performed on the dataset. Prior to the analysis, the data were standardized to ensure that all features were on a consistent scale. PCA is a dimensionality reduction technique that transforms the original set of features into a new set of uncorrelated variables known as principal components. These components capture different amounts of variance in the data, with each component representing a linear combination of the original features. By analyzing the explained variance ratio for each principal component, we can determine which nutrition factors contribute the most to the overall variability in the dataset. The principal components with higher explained variance ratios are indicative of nutrition factors that exhibit significant variations and are more likely to have a substantial impact on the subsequent clustering analysis. We consider all seven nutritional factors to identify the principal nutritional components that shows higher variant. In following table shows the basic properties of our original dataset before we preprocess data (Table 2 - Properties of Original dataset).

Table 2 - Properties of Original dataset

Source	Number of recipes	Number of Columns	Number of Ratings	Number of comments	Number of Recipes with comments
Valio Oy	4833	57	14081	24630	3060

To comprehensively evaluate the sentiment expressed in the dataset, we harnessed the powerful Vader Python sentiment analysis library. By applying this library to each comment, we obtained sentiment into three properties: neutral, positive, and negative. To derive a holistic sentiment representation for each comment, we calculated the average sentiment based on these properties.

However, we encountered a challenge with the Afinn Python library, which did not inherently support the Finnish language. To overcome this limitation, we established a local environment where we could configure and utilize the Afinn sentiment analyzer. Although Afinn library already offered support for several Nordic languages, such as Danish and Norwegian, it did not include support for Finnish. By setting up the Afinn sentiment analyzer locally, we successfully harnessed its sentiment analysis capabilities and applied it to retrieve sentiment scores for each comment in the dataset.

In a manner that we followed with our approach using the Vader library, we calculated the average sentiment for each comment by considering the sentiment scores obtained from Afinn. This process allowed us to capture a comprehensive sentiment overview of the comments associated with the recipes. By aggregating the sentiment scores and calculating their average, we derived a final sentiment value for each recipe. This value served as a measure of the overall sentiment associated with that recipe.

By incorporating an attributed graph representation, we aimed to enhance the mere connections between nodes and edges. Attributes served to provide supplementary context, properties, or characteristics associated with the nodes or edges, thereby enriching the overall data representation. This inclusion of attributes allowed us to capture a more comprehensive and nuanced understanding of the underlying data.

To achieve a visually informative graph visualization, we employed the Kamada-Kawai layout algorithm. This algorithm optimized the arrangement and positioning of the graph elements, ensuring that related nodes and edges were positioned in close proximity to one another. By leveraging the Kamada-Kawai layout algorithm, we try to enhance the clarity and readability of the visual representation.

Overall, the utilization of attributes within the graph representation enabled us to encapsulate a broader range of information and context, providing a more detailed and comprehensive depiction of the data. Additionally, the application of the Kamada-Kawai layout algorithm contributed to an aesthetically pleasing and intelligible visualization, facilitating effective exploration and understanding of the graph structure and its associated attributes and we used the Gephi tool for better graphical representation.

After that to obtain meaningful insights from the data, we employed the K-means clustering algorithm. This technique facilitated the identification of underlying patterns or trends that may not be readily apparent. By clustering the data based on similarity, we aimed to uncover valuable information in various domains such as image analysis, customer segmentation, and anomaly detection. Specifically, we focused on exploring patterns related to the nutritional content of the recipes.

To determine the optimal number of clusters for the data, we utilized the Silhouette coefficient, which is a statistical measure employed to evaluate the quality of clustering results. This coefficient provided a quantitative assessment of how well each data point fit within its assigned cluster, aiding in the determination of the appropriate number of clusters.

In addition to K-means clustering, we leveraged Spectral Clustering to identify non-linear boundaries within the dataset. Spectral clustering is particularly useful when dealing with complex data structures and enables the discovery of clusters that exhibit non-linear relationships. By employing this technique, we aimed to uncover intricate associations and patterns in our food recipe dataset, with a specific focus on nutritional factors. Overall, our aim is to reveal hidden patterns using K-means clustering and Spectral Clustering, detect non-linear relationships, and gain insights into the nutritional aspects of the recipes.

In the subsequent step, we focused on examining the correlations and coefficients among different variables within the dataset, specifically analyzing the relationship between nutritional values such as energy, protein, fat, and user-related metrics like user rating, user sentiment, and the number of comments.

As the available dataset did not provide sufficient user data and ratings, finally we made the decision to develop a content-based recommendation system. To implement this, we began by constructing a cosine similarity matrix based on the nutritional values of the recipes. Cosine similarity is a commonly used measure to determine the similarity between two vectors, disregarding their magnitudes. In our study, the cosine similarity matrix allowed us to identify recipes that were most similar to each other based on their nutritional content. This information served as the foundation for providing recommendations to users. Furthermore, we calculated a Food Standards Agency (FSA) factor for each recipe, considering the nutritional factors. The FSA factor played a critical role in determining the final recommendations provided to users. This content-based recommendation approach allowed us to provide users with personalized recommendations by considering the similarity between nutritional profiles and incorporating the health factor.

3.3 Concepts

In this section, we provide an overview of the fundamental methods that form the basis of our thesis research. We will delve into the concepts and techniques related to attributed graphs, sentiment analysis, Health Factors, Clustering Methods, and social media aspects. These methods are integral to our research and play a crucial role in achieving our research objectives. By exploring and analyzing these topics in detail, we aim to gain a deeper understanding of their applications and implications in the context of our thesis.

3.3.1 Attributed Graph

An attributed graph is a data structure that describes entities and their relationships, with attributes or qualities associated with both entities and relationships. It is an effective technique for modeling complicated systems and gathering detailed information about entities and their relationships.

Entities are represented as nodes in an attributed graph, and connections are represented as edges linking the nodes. Each node and edge can be associated with one or more characteristics, which offer extra information about the underlying object or connection. Depending on the nature of the data being represented, these qualities might be numerical, categorized, or textual.

Attributes associated with nodes in an attributed graph can specify a variety of properties of the entities they represent. In a social network graph, for example, nodes represent persons, and characteristics associated with each node might contain demographic information such as age, gender, or location. Nodes in a citation network can represent academic publications, with features such as title, authors, publishing venue, and citation count.

Similarly, in an attributed graph, attributes linked with edges capture information about the relationships between items. In a collaboration network, for example, edges may reflect researcher co-authorship, and attributes associated with each connection may include the number of co-authored articles or the strength of the cooperation.

Attributed graphs are utilized in a wide range of areas and applications. They have applications in social network analysis, recommendation systems, bioinformatics, knowledge graphs, and network science. Attributed graphs provide more thorough and nuanced study of complicated systems by integrating characteristics..

Finally, attributed graphs provide an expressive and rich framework for expressing and evaluating complicated systems. They offer a more complete knowledge of items and connections by adding qualities, hence facilitating diverse activities in many areas. To fully realize the potential of attributed graphs in real-world applications, however, difficulties relating to data sparsity, scalability, privacy, and security must be overcome [15] [16] [17].

3.3.2 Health Measurement of the Foods

We utilized the internationally recognized UK Food Standard Agency (FSA) "traffic light" system as a measure of the healthiness of a recipe or meal. This system, widely used in food labeling, provides a straightforward visual representation of a food's nutritional content and its overall healthiness.

In our study, we employed the FSA score, also known as the "traffic light" system, to assess the healthiness of the recipes in our dataset. The scoring methodology, as outlined in Table 3, takes into account the scaling of three macronutrients: fats, saturated fats, and salt. These macronutrients are categorized into three color-coded levels, ranging from green (healthy) to red (unhealthy).

To calculate the FSA score for each recipe, we followed the approach of previous studies and considered the available nutrient information, excluding sugar content due to data limitations. Based on the nutrient content of fats, saturated fats, and salt, we assigned scores of 1, 2, or 3 to represent the green, amber, and red categories, respectively. The total FSA score for each recipe ranged from 3 to 9, with a lower score indicating a healthier recipe and a higher score indicating a less healthy recipe(Table 3).

By incorporating the FSA score into our analysis, we were able to evaluate the healthiness of recipes and consider this factor when making recommendations to users. This approach allowed us to prioritize healthier options and provide users with a more informed understanding of the nutritional content of the recommended recipes.

Table 3

Dietary Factor	Low (1)	Medium (2)	High (3)
Fat	3% or less	3 - 17.5 %	17.5% or more
Saturated fats	1.5% or less	1.5 - 5 %	5% or more
Salt	0.3% or less	0.3 - 1.5 %	1.5% or more

3.3.3 WHO Health Factors and FSA Health Factors

The World Health Organization (WHO) emphasizes the importance of food in health promotion and maintenance. The World Health Organization lists many significant food-related health aspects that contribute to general well-being and illness prevention. These features of food consumption include nutritional value, food safety, and dietary habits. Understanding and addressing these variables is critical for people, communities, and governments to make educated decisions and put in place successful public health measures.

- Adequate Nutrition: Adequate nutrition is essential for optimum health. The World Health Organization highlights the necessity of eating a well-balanced diet that includes all important elements such as carbs, proteins, fats, vitamins,

and minerals. A varied and well-balanced diet helps to avoid malnutrition, promotes healthy growth and development, and lowers the risk of chronic illnesses.

- Nutritional content: The nutritional content of food is important in determining its health impact. WHO emphasizes the need of limiting unhealthy fats, added sugars, and excessive salt consumption, as they are connected to an increased risk of obesity, cardiovascular disease, diabetes, and other noncommunicable diseases (NCDs). An increased diet of fruits, vegetables, whole grains, and lean meats, on the other hand, can lead to improved health outcomes.
- Dietary Patterns: The World Health Organization acknowledges that overall dietary patterns have a substantial influence on health outcomes. It promotes the adoption of culturally acceptable, sustained healthy eating behaviors that enhance long-term well-being. The Mediterranean diet, which emphasizes fruits, vegetables, whole grains, legumes, and healthy fats while limiting processed foods, sugary drinks, and red meat, is an example of a healthy dietary pattern.
- Food Policies and Governance: Efficient food policies and governance are critical for providing an enabling environment that promotes healthy food choices while protecting public health. WHO urges governments to create evidence-based policies, laws, and initiatives to promote healthy eating habits, enhance food safety, and address food-related issues. Collaboration among many stakeholders, such as government agencies, industry, civil society, and academia, is critical for formulating comprehensive and effective food policy.

These WHO food-related health variables provide a comprehensive framework for comprehending the multidimensional nature of nutrition, food intake, and their influence on health. Individuals, communities, and politicians may work together to create healthier food environments, improve dietary habits, and eventually improve the general well-being of populations globally by addressing these aspects holistically [18].

In the United Kingdom, the Food Standards Agency (FSA) is responsible for guaranteeing the safety and quality of food ingested by the general population. The FSA is in charge of establishing and implementing food standards, as well as conducting research and offering advice in order to promote good health and safeguard consumers from food-related dangers.

The FSA acknowledges the important influence that dietary choices may have on an individual's well-being when it comes to food-related health problems. A nutritious diet is critical for sustaining good health, avoiding chronic illnesses, and boosting general well-being. The FSA highlights the following food-related health factors:

- Balanced Nutrition: The Food Standards Agency (FSA) acknowledges the critical value of eating a well-balanced diet that provides the body with all of the nutrients it requires to operate properly. A well-balanced diet includes a wide range of dietary categories such as carbs, proteins, fats, vitamins, and minerals. Carbohydrates are the body's principal source of energy, but proteins are required for tissue development, repair, and maintenance. Fats give energy, and insulation, and assist in the absorption of fat-soluble vitamins when ingested in moderation. Furthermore, vitamins and minerals are essential for

metabolic functions, immunity, and general health. The FSA hopes to enhance individuals' growth, development, and general well-being by encouraging a balanced diet.

- Portion management: The FSA emphasizes portion management as a critical component of maintaining a healthy eating pattern. It is easy to lose sight of portion sizes and consume more calories than the body requires in today's food-centric culture. Overeating can result in weight gain, which raises the risk of a variety of health problems, including obesity, diabetes, and cardiovascular disease. The FSA hopes to empower individuals to manage their energy intake properly, maintain a healthy weight, and lower their risk of acquiring chronic diseases by encouraging them to be conscious of portion sizes. Paying attention to serving sizes, listening to internal indications of hunger and fullness, and avoiding large quantities often given in restaurants and fast-food places are all part of practicing portion control.
- Nutrient Density: Promoting the consumption of nutrient-dense foods is another significant emphasis area for the FSA. Nutrient-dense foods have a high concentration of vital nutrients in comparison to their calorie value. Fruits, vegetables, whole grains, lean proteins, and low-fat dairy products are examples of such foods. These foods are high in vitamins, minerals, and antioxidants, all of which are essential for good health and well-being. Fruits and vegetables are high in vitamins A, C, and E, as well as minerals such as potassium and folate. Whole grains are high in fiber, B vitamins, and minerals like magnesium and iron. Lean proteins, such as poultry, fish, lentils, and tofu, are high in essential amino acids, which are required for tissue growth and repair. Low-fat dairy products offer calcium, vitamin D, and protein. By encouraging nutrient-dense options, the FSA hopes to guarantee that people get the nutrients they need while also controlling their overall calorie intake.
- Salt Reduction: The Food Standards Agency (FSA) acknowledges that excessive salt consumption is closely connected to high blood pressure, a major risk factor for cardiovascular disease. The FSA advises people to minimize their salt intake in order to improve their cardiovascular health. This can be accomplished by selecting low-sodium options, such as reduced-salt bread, soups, and sauces. Individuals are also encouraged to restrict their usage of salt when cooking and at the table. Individuals may make educated decisions and choose goods with reduced salt levels by reviewing food labels for sodium content. The FSA's emphasis on salt reduction seeks to enhance public health outcomes by lowering the prevalence of high blood pressure and the burden of cardiovascular illnesses.
- Fat Quality: The FSA emphasizes the necessity of eating healthy fats while reducing saturated and trans fats. Healthy fats, such as monounsaturated and polyunsaturated fats, are essential for overall health, particularly heart health. Nuts, peanuts, avocados, and fatty seafood like salmon and mackerel are good sources of healthful fats. These fats are advantageous because they aid in the reduction of dangerous cholesterol levels and the promotion of cardiovascular health. Saturated fats, which are often found in fatty meats, butter, and full-fat

dairy products, on the other hand, can raise the risk of cardiovascular disease when ingested in excess. Similarly, trans fats, which are commonly present in processed and fried meals, are known to be harmful to heart health. The FSA recommends that people restrict their consumption of saturated and trans fats and instead choose healthier fat sources. The FSA strives to enhance cardiovascular health outcomes and lower the burden of heart disease in the population by boosting the consumption of healthy fats and minimizing the consumption of saturated and trans fats.

By promoting these food-related health issues, the FSA hopes to enable consumers to make educated decisions, change their dietary habits, and live healthier lives. The organization collaborates with the government, businesses, and consumers to create a food environment in the United Kingdom that supports and promotes good health [19].

3.3.4 Clustering Methods

Clustering is a fundamental approach in data analysis and machine learning that involves grouping together similar items based on their intrinsic properties. Pattern recognition, picture analysis, social network analysis, marketing, and recommendation systems are just a few of the applications. Clustering approaches seek to uncover hidden structures and correlations in data, delivering useful insights and easing decision-making.

We will go into the realm of clustering methods in this complete introduction, studying its ideas, types, algorithms, and applications. We will explore both old and new clustering approaches, emphasizing their advantages, disadvantages, and major implementation issues. So, we separated it into subtopics and discussed them.

Introduction: Clustering is an unsupervised learning problem, which means that no pre-labeled data or goal outputs are required for training. The purpose is to divide a dataset into groups or clusters, with objects within the same cluster showing high similarity and objects in other clusters showing dissimilarity. The essential premise is that items inside the same cluster have certain characteristics and may be considered a coherent subset.

Clustering rules: Specific rules regulate the process of grouping related items in clustering algorithms. Some fundamental principles are as follows:

a. Similarity or Distance: Clustering is based on determining how similar or different entities are. To assess the dissimilarity of data points, several distance metrics such as Euclidean distance, Manhattan distance, or cosine similarity are utilized.

b. Cluster Separation: Clusters should be distinct and well-separated from one another, with as little overlap as possible. Inter-cluster distance measurements are frequently used to analyze cluster separation.

c. Cluster Coherence: Objects in the same cluster should be internally coherent, with high similarity and low within-cluster variation. Intra-cluster distance measurements are commonly used to assess internal cluster coherence.

Algorithms for Clustering: Various algorithms have been developed to implement various clustering strategies. Let's take a quick look at some prominent clustering algorithms:

a. K-means: K-means is an iterative clustering technique that is commonly used for clustering data points. The sum of squared distances between data points and cluster centroids is minimized. The process in K-means begins by randomly initializing cluster centroids and then allocates each data point to the nearest centroid based on distance metrics, often Euclidean distance. The centroids are updated following the assignment stage by determining the mean of the data points inside each cluster. This method is repeated until the centroids no longer vary appreciably or the maximum number of iterations is achieved. K-means is extensively used in numerous disciplines for exploratory data analysis and pattern identification tasks because it is good at splitting data into compact, spherical clusters.
b. Spectral Clustering: Using graph theory and eigenvalue analysis, spectral clustering splits data points into groups. It provides an alternative to typical clustering approaches. The process of spectral clustering begins with the construction of a similarity graph, in which each data point is represented as a node and the edges record the pairwise similarity between the data points. After that, the graph is translated into a lower-dimensional space using eigenvalue analysis on the graph Laplacian matrix. This stage decreases the dimensionality of the data and aids in revealing its underlying structure. Finally, the altered data is subjected to a clustering algorithm, which groups the points into clusters. Spectral clustering can capture complex relationships and nonlinear structures in the data, making it useful for data exploration and clustering tasks where traditional methods might struggle.

Finally, clustering approaches are effective for detecting patterns, grouping related data points, and getting insights from large datasets. Practitioners may efficiently use clustering approaches to many domains and extract meaningful information from their data by knowing the concepts, types, algorithms, and applications of clustering. However, for relevant and trustworthy clustering findings, thorough consideration of data pretreatment, algorithm selection, assessment methodologies, and domain understanding is required [20] [21] [22] [23].

3.3.5 Social Media & Social Network Analysis

Online platforms and websites that allow users to produce and share information, communicate with others, and engage in virtual communities are referred

to as social media. It has changed the way people communicate, share information, and interact with one another. Social media platforms have become a vital element of modern life, giving numerous possibilities for individuals, businesses, and organizations to communicate, cooperate, and share information.

Social network analysis (SNA) is the study of the linkages and interactions between persons or entities in a social network. It examines the structure, patterns, and dynamics of social interactions in order to gain insight into how information moves, influence is exercised, and communities emerge and change.

SNA is especially important in social media because it helps scholars and practitioners to grasp the complex networks of relationships established by online platforms. It gives useful information on user habits, social impact, knowledge diffusion, and community dynamics. SNA uncovers hidden structures and identifies significant players or influential persons inside a social network by evaluating the relationships between users, their interactions, and the patterns of information exchange.

To summarize, social media and social network analysis have transformed the way people and organizations connect, communicate, and exchange information. SNA gives useful insights into the complex network architecture, social behaviors, and information dynamics that exist on social media platforms. Researchers and practitioners may get a deeper knowledge of social phenomena, devise successful tactics, and make informed judgments in a variety of fields by studying user connections and interactions [24] [25] [26].

4 RESULTS & DISCUSSIONS

In this chapter, we present the detailed results and comprehensive analysis of the experimental phases conducted in our thesis. By diligently implementing the concepts and methodology outlined in the previous chapter, we carefully examine the outcomes obtained from our research endeavors. Through evaluation and data processing, we ensure the reliability and relevance of the findings to our research objectives. The analysis provides meaningful insights and contributes to the existing body of knowledge in the field.

4.1 Step 1 – Graphical Analysis

In our analysis, we sought to identify relationships and clusters within the dataset by generating a graphical representation using the Kamada-Kawai layout algorithm

However, upon visual inspection of the graph, we were unable to discern any clear clusters or distinct patterns. The graph generated using the Kamada-Kawai layout algorithm did not provide a clear visual representation of the underlying relationships and groupings within the data.

To further investigate and gain deeper insights into the dataset, we utilized the Gephi tool to generate an alternative graph representation (Figure 1- Generated Graph Using Gephi). The graph generated through Gephi allowed for a more comprehensive exploration of the data structure, enabling us to identify four distinct clusters with greater clarity. These clusters represented meaningful groupings and relationships among the data points. We used this information to perform the clustering techniques.

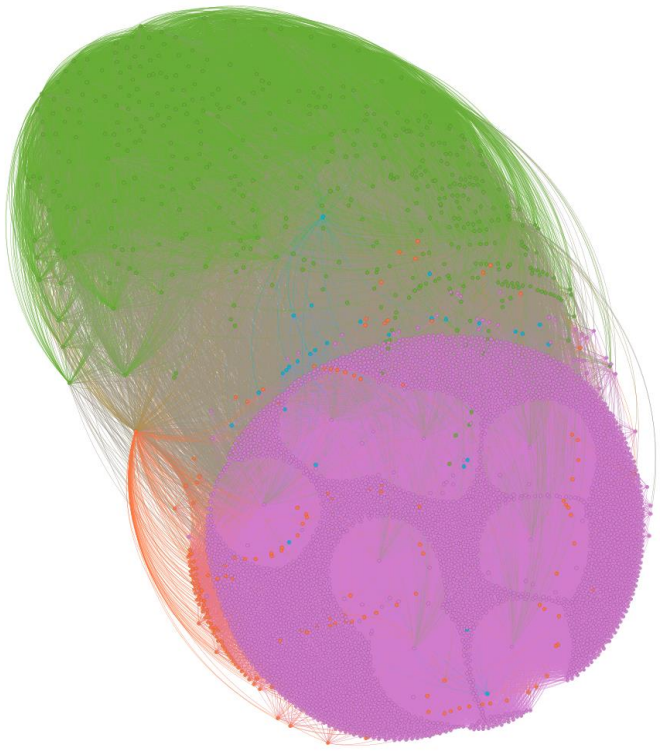


Figure 1- Generated Graph Using Gephi

4.2 Step 2 – Sentiment Analysis

In the sentiment analysis phase of our research, we utilized two libraries to account for the language requirements of our dataset. Initially, we had 4,833 records in our dataset, but after preprocessing, we were left with 2,587 recipes that had user comments available for analysis. To determine the sentiment of each comment, we employed the AFINN sentiment analyzer, which provided sentiment analysis based on the original language of the comments. Additionally, we used the VADER sentiment analysis by translating each comment from Finnish to English using Google Translate.

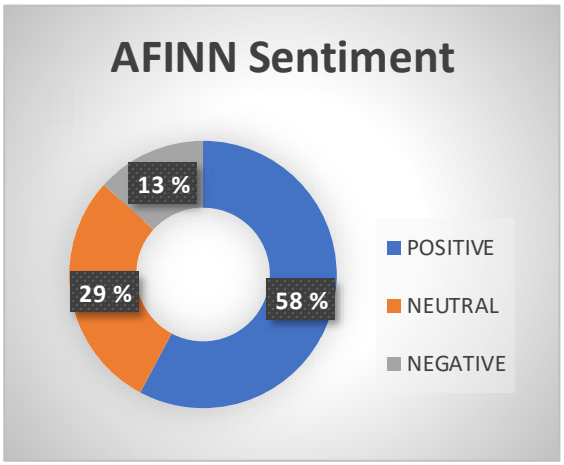


Figure 3 - AFINN Sentiment

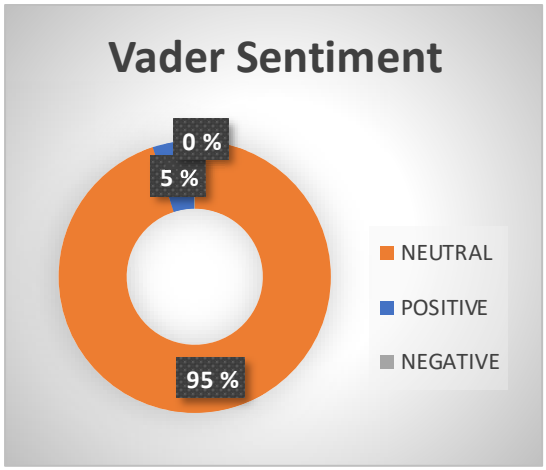


Figure 2 - Vader Sentiment

Table 4 - Summary of AFINN Sentiment

AFINN Sentiment	Count
POSITIVE	1495
NEUTRAL	748
NEGATIVE	344
Grand Total	2587

Table 5 - Summary of Vader Sentiment

Vader Sentiment	Count
NEUTRAL	2458
POSITIVE	125
NEGATIVE	4
Grand Total	2587

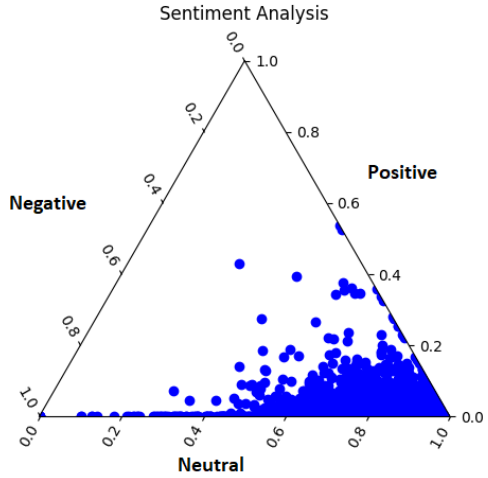


Figure 4 - Triangular representation Vader Sentiment

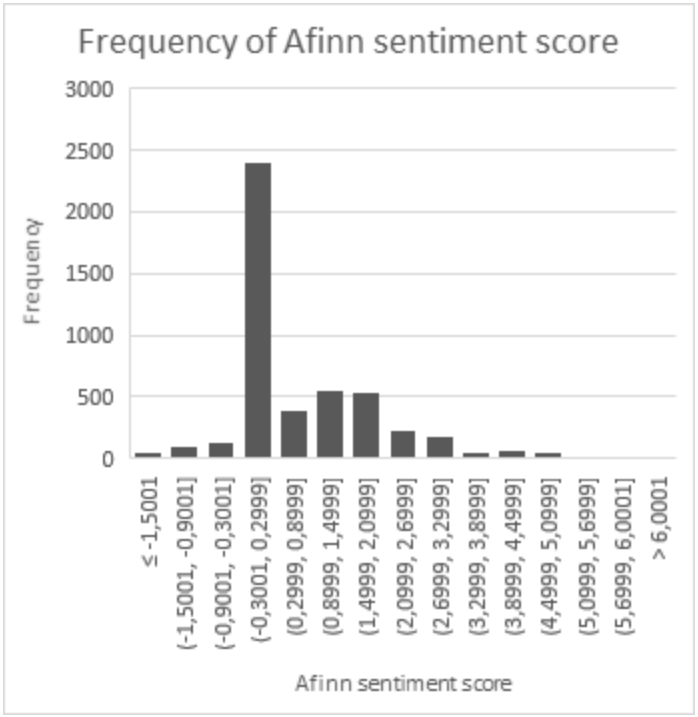


Figure 5 - AFINN sentiment score

We utilize the AFINN sentiment values to categorize sentiments into three distinct categories: Negative, Neutral, and Positive. To ensure consistency and clarity in our classification, we have established specific thresholds. Sentiments with a value below -0.5 are considered Negative, indicating a predominantly negative sentiment. Sentiments falling within the range of -0.1 to +0.1 are classified as Neutral, signifying

a lack of strong positive or negative sentiment. Finally, sentiments with a value greater than +0.5 are categorized as Positive, representing a generally positive sentiment. By implementing this categorization scheme, we can effectively differentiate and evaluate sentiments based on the provided AFINN sentiment values.

Upon comparing the results of the two analyses, we observed a significant difference between the outcomes generated by AFINN and VADER. These two methods yielded contrasting results, indicating a discrepancy in sentiment classification. However, in order to proceed with our analysis, we relied on the sentiment results obtained from the AFINN analysis, as it considered the original words and context of the comments. It became apparent that the translation of comments had a considerable impact, as many comments did not yield meaningful outputs in English.

After pre-processing data, we have 2587 recipes that contain comments that can be used for further analysis out of 3060. According to the AFINN sentiment analysis(Figure 3), 58% of the user comments were classified as positive, amounting to 1,495 out of 2,587. Approximately 29% of the comments were deemed neutral, totaling 748 out of 2,587, while 13% were identified as negative, accounting for 344 out of 2,587, (a summary of results can be shown Table 4). Conversely, the sentiment analysis performed by VADER (Figure 2) categorized 95% of the comments as neutral, encompassing 2,458 out of 2,587, with only 5% classified as positive, equivalent to 129 out of 2,587. Interestingly, based on the VADER analysis, a mere 4 recipes out of 2,458 were determined to have negative sentiment (A summary of VADER sentiment can be shown

Table 5). A summarized AFINN and Vader sentiment analysis distribution can be shown in Figure 4 & Figure 5Figure 4 - Triangular representation Vader Sentiment

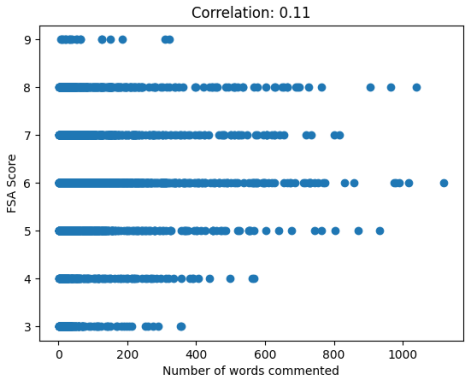
The significant contrast between the sentiment results obtained from AFINN and VADER suggests that the translation of comments may have altered the original meaning of the text. Consequently, we proceeded with the sentiment analysis based on the AFINN results, taking into account the sentiment derived from the original comments. We hypothesize that the observed differences can be attributed to the semantic nuances lost during the translation process.

Upon translating the comments into English, we observed a discrepancy between the intended meaning and the resulting translation. The translated version does not accurately convey the original intent and may lead to misunderstandings or misinterpretations. It is crucial to recognize that translations can sometimes alter the intended message, emphasizing the need for careful evaluation and accurate representation of the comments, the other fact is most of the users gave their comment and then ask many questions about recipe or the preparation methods. Therefore we suspect that VADER gave Neutral for many user comments (See below example)

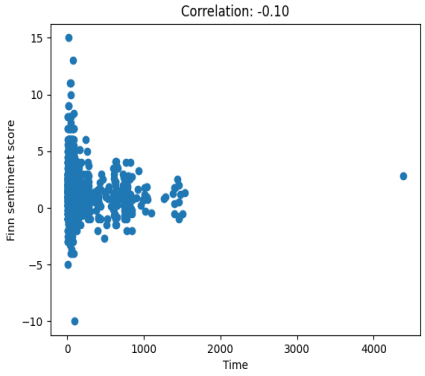
- Mahtava sitruunapommi, koukuttava maku! - Awesome lemon bomb
- mutta ensikertalainen menee tällä ohjeella metsään - a first-timer will go into the forest with this instruction

4.3 Step 3 – Correlation Analysis

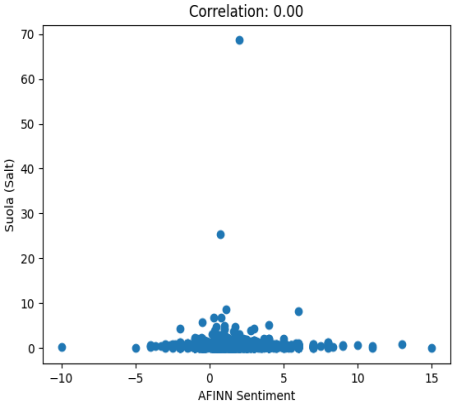
In our study, we conducted a thorough statistical analysis to investigate the potential correlations between user feedback and the content of food recipes. As well as we investigate the correlations between number of words commented by users per each recipe and nutritional entities. By this we hoped to identify correlation between the number of words against each nutritional factor. By finding the correlation between the number of words and nutritional factors we hoped to identify any relationship between users commenting behaviors againts the nutrinal factors. We also took into account additional factors such as preparation time and steps to explore any possible relationships. However, our analysis did not reveal any strong associations between these variables. We calculate correlation between nutritional components and sentiment of the comments as well as the number of words contained in each recipe.



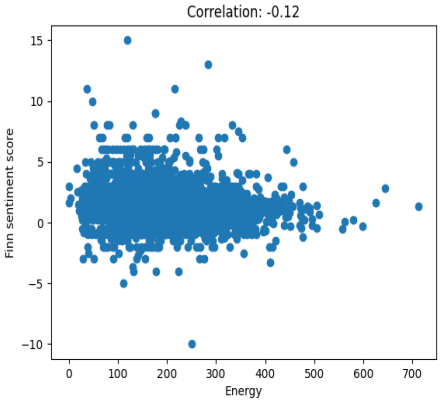
Graph 2 - FSA - Num of comm. words



Graph 3 - Preparation time and AFINN sentiment



Graph 5 - Salt and AFINN sentiment



Graph 4 - Energy and AFINN sentiment

Table 6 - Summary of correlation

We have attached a few graphs that shows correlation between several factors. In Graph 2 it shows the correlation between FSA factor and the number of commented words by the user for each recipe. In Graph 3 it shows the correlation between preparation time and AFINN sentiment score. Graph 5 represents the correlation between Salt level and the AFINN sentiment score. In Graph 4 it exhibits the correlation between AFINN sentiment score and Energy level. These are few graphs we attached and the Table 6 shows the all the correlational values between each nutritional factor and AFINN sentiment, number of words commented by users for each recipe. Despite considering various aspects of the recipes and user feedback, we did not find substantial evidence to suggest a significant correlation. This suggests that factors beyond the recipe content alone contribute to user feedback, making it a complex and multifaceted phenomenon.

	Ene rgyy	Prot ein	Carbo h ydrate	Fat	Saturated Fat	Dietar Fiber	Salt	FSA Factor
AFINN Sentiment	-0.12	-0.03	-0.09	-0.09	-0.09	-0.01	0.00	-0.08
Num Of Comm words	0.14	0.00	0.13	0.11	0.11	0.00	-0.02	0.11

The inclusion of preparation time and steps was an attempt to capture additional dimensions that might influence user feedback. However, even with these factors considered, no strong relationships emerged. This indicates that user feedback is influenced by various subjective factors that extend beyond the recipe's content and preparation details. These findings highlight the intricate nature of user preferences and the challenges in establishing direct correlations between recipe content and feedback (RQ1).

4.4 Step 4 – Clustering

In our endeavor to unravel the underlying structure of the dataset, we adopted a multi-faceted approach that encompassed both graph analysis and clustering techniques. Our objective was to determine the optimal number of clusters and explore the dataset's inherent patterns and relationships.

Through the Silhouette analysis, we examined the clustering outcomes and evaluated how well the data points were assigned to their respective clusters. This analysis provided us with valuable insights into the structure and cohesion of the dataset, allowing us to make informed decisions about the optimal number of clusters.

In order to analyze the nutritional components of our dataset and facilitate subsequent clustering techniques, we conducted a Principal Component Analysis (Table 7). Looking at the cumulative explained variance array [0.44046379, 0.62722796, 0.78347269], we can see that the first principal component (PC1) explains 44.05% of the total variance, the first two components (PC1 and PC2) explain 62.72% of the total variance, and all three components (PC1, PC2, and PC3) explain 78.35% of the total variance. Through PCA, we identified the principal components that exhibit significant variance among the nutritional factors. Specifically, our analysis revealed that Energy, Carbohydrate, and Fat demonstrated substantial variability. These identified nutritional components served as the basis for our subsequent analyses, including Silhouette analysis and Spectral Clustering. By utilizing the PCA results and focusing on these key nutritional components, we aimed to gain insights into the clustering patterns and relationships within our dataset, ultimately contributing to a comprehensive understanding of the underlying nutritional characteristics.

In this particular case, after careful evaluation of the Silhouette coefficients for different cluster configurations, we determined that the dataset was most effectively divided into 2 clusters. The Silhouette coefficient, which measures the similarity of data points within their assigned clusters compared to other clusters, played a crucial role in this determination.

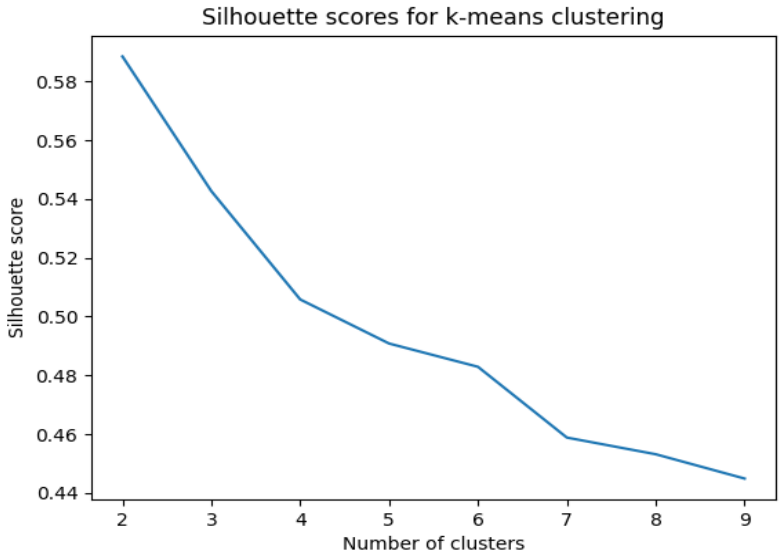
$$s(i) = \frac{b(i) - a(i)}{k\max\{a(i), b(i)\}}$$

where:

- $s(i)$ is the silhouette coefficient for data point i.
- $a(i)$ is the average dissimilarity between data point i and all other data points within the same cluster.
- $b(i)$ is the average dissimilarity between data point i and all data points in the nearest neighboring cluster (the cluster to which i does not belong).

Table 7 - PCA results

	PC1	PC2	PC3	PC4	PC5	PC6	PC7
Energia	0.559985	-0.06333	0.025685	0.012013	-0.17628	0.27829	0.757022
Proteiini	0.148767	0.51371	0.416405	-0.62607	-0.35124	-0.11325	-0.11142
Hiilihydraatit (Carbohydrates)	0.354919	-0.50566	0.149338	0.209047	-0.60825	-0.15674	-0.39724
Rasva (Fat)	0.51952	0.19825	-0.18338	-0.00363	0.276098	0.569537	-0.50651
Tyydyttynyt rasva (Saturated fat)	0.490106	0.206345	-0.30282	0.074116	0.258849	-0.74365	-0.00253
Ravintokuitu (Dietary fiber)	0.167116	-0.43812	0.630692	-0.20561	0.576344	-0.08809	-0.01182
Suola (Salt)	0.016892	0.450072	0.529668	0.718616	0.01348	-0.00335	0.000155



Graph 6 - Silhouette scores

By identifying the optimal number of clusters through Silhouette analysis, we ensured the resulting number of clusters. This knowledge serves as a foundation for further exploration and analysis, enabling us to delve deeper into the distinct characteristics and patterns exhibited by each cluster within the dataset. Utilizing the insights gained from the Silhouette analysis, we proceeded to employ the K-means clustering algorithm to partition the dataset into the determined number of clusters (Figure 6).

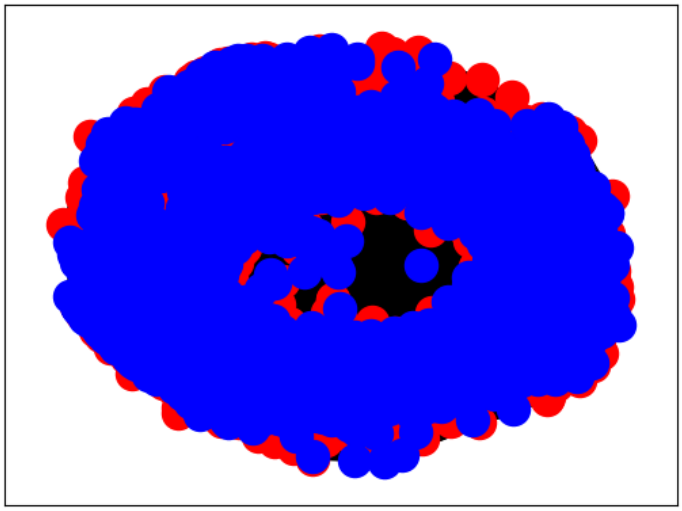


Figure 6- K-Means clustering

Since, we can clearly identify there are four clusters in the attributed graph, we expanded our analysis by applying spectral clustering, a powerful algorithm that excels at identifying clusters with non-linear boundaries (Figure 7). We used the principal components that we learned by doing the PCA to cluster our recipes. By utilizing this technique, we were able to unravel the intricate relationships and non-linear associations within the dataset. The spectral clustering process yielded a partitioning of the data into four distinct clusters, unveiling previously unrecognized patterns and dependencies.

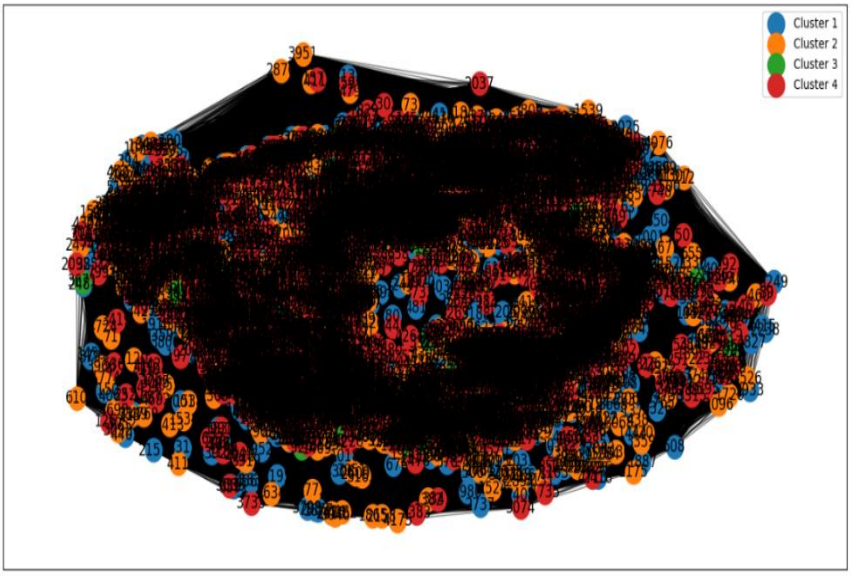
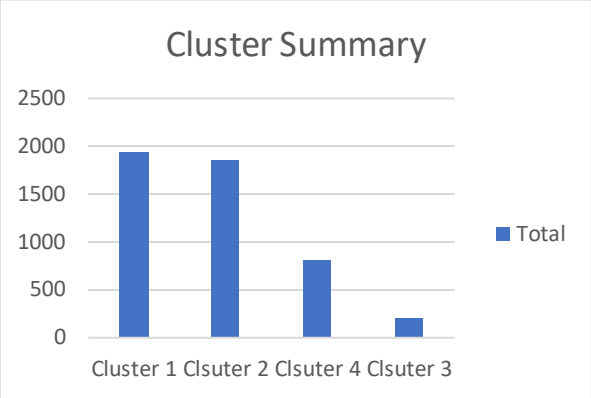


Figure 7 - Spectral Clustering graph

Following the clustering of our dataset, we conducted a comprehensive statistical analysis on each cluster, focusing on nutritional values, health factors, sentiment from user feedback, ratings, and the length of comments. Our dataset comprised four clusters, with Cluster 1 containing a highest number of recipes compared to the other clusters, accounting for 1950 out of 4,833 recipes. Cluster 2 consisted of 1859 recipes, Cluster 3 had 214 recipes, and Cluster 4 contains 810 recipes.

To perform our analysis, we considered several nutritional factors, including Energy, Protein, Carbohydrate, Fat, Saturated Fat, Dietary Fiber, and Salt. Additionally, we incorporated AFINN Sentiment, Vader positive and neutral sentiment values, the number of comments, the number of commented words, and the rating as user feedback indicators.



Cluster Class	Count
Cluster 1	1950
Cluster 2	1859
Cluster 4	810
Cluster 3	214
Grand Total	4833

Table 8- Average of each nutrition component

Cluster	Energy	Protein	Carbohydrate	Fat	Saturate Fat	Dietar Fiber	Salt	FSA Factor
Cluster 1	282.38	6.3	27.61	15.8	7.59	1.53	0.5	6.57
Cluster 2	142.63	6.6	9.33	8.42	3.80	1.05	0.6	5.85
Cluster 3	48.65	1.8	7.99	0.5	0.0	0.58	0.1	3.15
Cluster 4	79.55	4.7	5.93	3.76	1.53	0.95	0.6	4.98

Cluster	AFINN Sentiment	Vader Sentiment Positive	Vader Sentiment Neutral	No of Comments	Number of commented words	Rating
Cluster 1	1.462	0.188	0.762	5.379	86.894	4.36
Cluster 2	1.292	0.185	0.759	8.738	108.209	4.38
Cluster 3	1.476	0.187	0.752	8.325	108.901	4.37
Cluster 4	1.968	0.205	0.748	2.538	61.230	3.77

Among the clusters of recipes analyzed, Cluster 1 stood out with the highest energy levels, indicating a significant contrast compared to the other clusters. This suggests that the recipes in Cluster 1 are generally more calorie-dense and potentially offer more substantial meals. However, when considering the protein content, Cluster

1 and Cluster 2 displayed similar levels, with Cluster 1 at 6.39 and Cluster 2 at 6.63. In contrast, Cluster 3 showed a notably lower protein level, measuring only 1.89. This indicates that recipes in Cluster 3 may provide comparatively lower amounts of protein.

Looking at the carbohydrate levels, Cluster 1 exhibited a substantial contrast compared to the other clusters. The carbohydrate level in Cluster 1 was 27.61, significantly higher than the second largest value of 9.33 found in the other clusters. This difference suggests that recipes in Cluster 1 may contain significantly more carbohydrates, which could make them a suitable choice for individuals seeking higher carbohydrate intake.

In terms of fat content, Cluster 3 showed a significant contrast when considering both fat and saturated fat levels. The recipes in Cluster 3 had the lowest fat content, with a value of 0.5, and completely lacked saturated fat (0.0). On the other hand, Cluster 1 displayed the highest levels of both fat and saturated fat. This suggests that recipes in Cluster 1 may have a higher fat content and can contribute to a higher intake of saturated fat compared to the other clusters.

Furthermore, when considering the FSA (Food Standards Agency) factor, Cluster 1 had the highest FSA value among all the clusters. This indicates that recipes in Cluster 1 may be associated with a higher FSA score, which implies a potentially higher level of protein, carbohydrate, and fat content. In contrast, Cluster 3 had the lowest FSA value, suggesting that recipes in this cluster may have a lower overall FSA score, indicating a healthier composition in terms of salt, sugar, and fat content. (Table 8- Average of each nutrition component).

However, a significant concern about the dataset is Cluster 1 and cluster 2 accounted for more than 78.8% of the entire dataset. This uneven distribution within the clusters led us to decide against considering these relationships when recommending recipes to users. Even though these findings provided valuable insights into the dataset, they were not supposing this suitable for our recipe recommendation system due to the imbalanced cluster distribution (RQ2).

4.1 Step 5 – Health-based Recommender

In our recommendation system, we have employed a content-based filtering approach due to the absence of user-related data. The system utilizes recipe nutritional factors, including Energy, Protein, Carbohydrate, Fat, Saturated Fat, Dietary Fiber, and Salt, to calculate the cosine similarity between each recipe. Additionally, we have incorporated a health factor calculation when determining the most similar food recipes.

In our recommendation system, we have employed a content-based filtering approach due to the absence of user-related data. The system utilizes recipe nutritional factors, including Energy, Protein, Carbohydrate, Fat, Saturated Fat, Dietary Fiber, and Salt, to calculate the cosine similarity between each recipe. Additionally, we have incorporated a health factor calculation when determining the most similar food recipes.

To enhance the system's performance and provide optimal results, we have set a maximum limit of 20 recommended recipes. Users have the flexibility to specify the desired number of recommendations through the provided API. Upon user request, the API returns the recommended recipes, along with the mean squared error, which serves as a valuable indicator for assessing the recommendation quality (RQ3).

Example Calculation –

We computed the Mean Squared Error (MSE) and similarities according to the given formulas. The k most similar foods were identified based on the cosine similarity, where k was set to 3. The MSE was calculated by considering the similarity between food items, and the health factor was determined based on the FSA factor. The Sum of MSE and Sum of health factor were aggregated to accumulate the values for further analysis. Suppose cosine similarity matrix is Table 9.

- **Cosine Similarity:**

Table 9- Sample cosine similarity matrix

	F1	F2	F3	F4	F5	F6
F1	1	0.5	0.4	0.3	0.7	0.6
F2	0.5	1	0.8	0.2	0.9	0.3
F3	0.3	0.8	1	0.4	0.8	0.2
F4	0.4	0.2	0.4	1	0.6	0.9
F5	0.7	0.9	0.8	0.6	1	0.7
F6	0.6	0.3	0.2	0.9	0.7	1

And health factor for each recipe as follow:

- **FSA Factor:**

	F1	F2	F3	F4	F5	F6
FSA factor	4	3	7	9	5	8

For recipe f_i Find k recipes which the multiply of the health factor of that food and similarity values of that food with f_i are maximum

$[f_i, f_i, f_i, \ldots f_k]$ (consider k = 3)

MSE=1- $\frac{\sum_{j=1}^{k}(sim\,(f_i,f_j)}{k}$

Health Factor:

Health=$\sum_{j=1}^{k}(health(f_j)$

SumMSE =MSE+SumMSE

SumHealth=SumHelath+Health

For example, a section of the cosine similarity matrix and FSA factor is provided. We showcased the computation of the content-based recommendation using the MSE and the average health factor. Below matrix is contains sample value of similarity between each food (F1, F2, ..)

- **Health factor: 1 - (FSA Factor - 3) / 6.0**

	F1	F2	F3	F4	F5	F6
Health factor	0.83333	1	0.33333	0	0.66666	0.16666

- Here, we consider k=3
- We have one loop for all foods:
- Food F1: Most similar food with food F1 based on cosine similarity are: [F5, F6, F2]
- MSE for F1: 1-$\frac{0.7+0.6+0.5}{3}$=0.4
- Average Health **for recommended food based on food F1:** $\frac{0.666667 + 0.166667 + 1}{3}$**= 0.60**

	F1	F2	F3	F4	F5	F6
Similarity × Health Factor	-	0.5	0.1333332	0	0.4666669	0.1000002

- **According to this scenario Healthy recommended foods are – F2, F5 and F3**

The mean squared error plays a crucial role in evaluating the accuracy of the recommendation system. By comparing the system's predictions to the actual nutritional values of the recommended recipes. By leveraging content-based filtering and considering nutritional factors, our recommendation system aims to deliver health-conscious recipe recommendations. The ability to customize the number of recommendations empowers users to receive suggestions that align with their preferences and dietary requirements.

In conclusion, our analysis of the food recommendation dataset revealed valuable insights regarding the correlation between nutritional factors, user feedback, and the design of a hybrid food recommender system. We found that the correlation between different nutritional factors and the popularity of foods in social media was not significant. This suggests that factors beyond nutrition alone play a crucial role in determining the popularity of food items on social media platforms (RQ1).

However, we found no strong correlations between nutritional factors, sentiment analysis, and user feedback. This indicates that user feedback is influenced by various subjective factors that extend beyond nutritional content, highlighting the complexity of understanding user preferences in the context of food recommendations. These findings emphasize the need to consider multiple factors and user-related data when designing accurate and effective food recommendation systems(RQ2).

We designed a hybrid food recommender system that incorporates nutrition content and a health factor. The system utilizes content-based filtering and calculates cosine similarity based on nutritional factors. Additionally, we integrated a health factor calculation to consider the overall healthiness of recommended foods. Incorporating user preferences and capturing them effectively remains a challenge, and future work should focus on incorporating user-related data to enhance the performance of the hybrid system.

Our study provides valuable insights into the correlation between nutritional factors, user feedback, and the design of a hybrid food recommender system. While the correlation between nutritional factors and popularity in social contribution was not significant, we identified dominant features influencing nutritional factors and designed a hybrid recommender system. However, challenges such as capturing user preferences, lack of user-related data, user feedback, and addressing imbalanced clusters need further investigation. This research contributes to the understanding of food datasets and lays the groundwork for future advancements in food recommendation systems.

5 SUMMARY

In this section, we present a comprehensive summary of our study, highlighting the steps we followed and the key findings we obtained. We began by analyzing the relationships between user feedback, recipe content, and additional factors such as preparation time and steps. We utilized statistical and clustering techniques to explore the dataset, uncover patterns, and identify distinct clusters. Furthermore, sentiment analysis was conducted on user comments to gain insights into the overall sentiment. Finally, we developed a recommendation system based on content-based filtering to offer personalized recipe suggestions. Let's now delve into the specific details of each step and the significant findings we derived from our investigation.

In our analysis, we initially attempted to identify relationships and clusters within the dataset using the Kamada-Kawai layout algorithm but found no clear patterns or clusters. To gain deeper insights, we turned to the Gephi tool, which allowed us to generate an alternative graph representation revealing four distinct clusters with greater clarity. These clusters represented meaningful groupings and relationships among the data points. Moving on to sentiment analysis, we used two libraries to analyze user comments, namely AFINN and VADER. AFINN provided sentiment analysis based on the original language of the comments, while VADER involved translating the comments from Finnish to English. Comparing the results, we found significant differences between AFINN and VADER, indicating a discrepancy in sentiment classification. Despite the differences, we relied on the AFINN sentiment analysis, considering the original words and context of the comments. AFINN categorized 58% of the comments as positive, 29% as neutral, and 13% as negative. On the other hand, VADER classified 95% of the comments as neutral, 5% as positive, and only 4 recipes as negative. We attribute these differences to the potential loss of semantic nuances during the translation process.

In our study, we conducted a statistical analysis to explore potential correlations between user feedback and the content of food recipes. Despite considering various factors, including preparation time and steps, we did not find strong associations between these variables. This suggests that user feedback is influenced by subjective factors beyond the recipe content alone, making it a complex phenomenon (RQ1).

To unravel the underlying structure of the dataset, we employed graph analysis and clustering techniques. Silhouette analysis helped us determine the optimal number of clusters, and we used the K-means and spectral clustering algorithms to partition the data. The analysis revealed four distinct clusters, with Cluster 3 being the largest, containing 86.5% of the dataset.

We performed a comprehensive statistical analysis on each cluster, focusing on nutritional values, sentiment from user feedback, ratings, and comment length. Cluster 1 exhibited the highest protein levels, Cluster 2 had the highest fat and saturated fat levels, and most recipes showed higher levels of dietary fiber. However, due to the imbalanced cluster distribution, we decided not to consider these relationships in our recipe recommendation system (RQ2).

In our recommendation system, we have implemented a content-based filtering approach since we lack user-related data. The system utilizes nutritional factors such as Energy, Protein, Carbohydrate, Fat, Saturated Fat, Dietary Fiber, and Salt from the recipes to calculate the cosine similarity between each recipe. This enables us to determine the similarity between recipes based on their nutritional content.

Additionally, we have integrated a health factor calculation to further refine the recommendations and prioritize healthier food options.

To ensure optimal performance and deliver relevant results, we have implemented a maximum limit of 20 recommended recipes. This allows users to specify the number of recommendations they desire through the provided API. Upon receiving a user request, the API generates the recommended recipes based on their nutritional similarity and health factors. Along with the recommended recipes, the API also provides the mean squared error as an important metric for assessing the quality of the recommendations. This metric helps evaluate the accuracy and effectiveness of the recommendation system (RQ3).

6 REFERENCES

[1] "Wikipedia. Recommender system.," Online. [Online]. Available: https://en.wikipedia.org/wiki/Recommender_system.

[2] Hadash, G., Shalom, O. S., & Osadchy, R., "Rank and rate. In Proceedings of the 12th ACM Conference on Recommender Systems," in *ACM*, 2018.

[3] A. Tugend, "Too many choices: A problem that can paralyze," *The New York Times,* p. 26, 2010.

[4] Lops, P., de Gemmis, M., & Semeraro, G., "Content-based recommender systems: State of the art and trends," in *Recommender Systems Handbook*, Springer, 2011, pp. 73-105.

[5] Adomavicius, G., & Tuzhilin, A., "Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions," IEEE Transactions on Knowledge and Data Engineering, 2005, pp. 734-749.

[6] B.O. Mbah, P.E. Eme and O.F. Ogbusu, "Effect of Cooking Methods (Boiling and Roasting) on Nutrients and Anti-nutrients Content of Moringa oleifera Seeds," in *Pakistan Journal of Nutrition*, Nsukka, 2012.

[7] Donghui Wanga, Yanchun Lianga, Dong Xua, Xiaoyue Fenga, Renchu Guan, "A content-based recommender system for computer science publications," in *ELSEVIER*, 2018.

[8] Christopher R. Aberger, "Recommender: An Analysis of Collaborative Filtering".

[9] J. Ben Schafer, Dan Frankowski, Jon Herlocker, Shilad Sen , "Collaborative Filtering Recommender Systems," in *Springer*.

[10] R. Burke, "Hybrid Recommender Systems: Survey and Experiments. User Modeling and User-Adapted Interaction," 2002.

[11] Balraj Kumar, Neeraj Sharma, "Approaches, Issues and Challenges in Recommender Systems: A Systematic Review," 2016.

[12] Katherine Harris-Lagoudakis, "Online shopping and the healthfulness of grocery purchases," Ames, Iowa.

[13] RACIEL YERA, AHMAD A. ALZAHRANI, LUIS MARTíNEZ, "A Food Recommender System Considering Nutritional Information and User Preferences," in *IEEE*.

[14] Raciel Yera Toledo, Ahmad A. Alzahrani, Luis Martínez, "A Food Recommender System Considering Nutritional Information and User Preferences," in *IEEE*, 2019.

[15] Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., & Yu, P. S., "A Comprehensive Survey on Graph Neural Networks," in *IEEE*, 2019.

[16] Lizi Liao, Xiangnan He, Hanwang Zhang, Tat-Seng Chua, "Attributed Social Network Embedding," in *TKDE*, 2018.

[17] Xu, K., Li, C., Tian, Y., Sonobe, T., Kawarabayashi, K. I., & Jegelka, S., "Representation learning on graphs with jumping knowledge networks," in *In Proceedings of the 32nd Conference on Neural Information Processing Systems (NeurIPS)*, 2018.

[18] "World Health Organization," [Online]. Available: https://www.who.int.

[19] "Food Standards Agency.," [Online]. Available: https://www.food.gov.uk/.

[20] Rui Xu, D. Wunsch, "Survey of clustering algorithms," 2005.

[21] A. K. Jain, "Data clustering: 50 years beyond K-means," 2010.

[22] Hastie, T., Tibshirani, R., & Friedman, J., "The Elements of Statistical Learning," 2009.

[23] Fei Wang, Hector-Hugo Franco-Penya, John D. Kelleher, John Pugh, "An Analysis of the Application of Simplified Silhouette to the Evaluation of k-means Clustering Validity," 2017.

[24] Kwak, H., Lee, C., Park, H., & Moon, S., "What Is Twitter, a Social Network or a News Media?," in *In Proceedings of the 19th international conference on World Wide Web*, 2010.

[25] Anna Charmantzi, Ioannis Antoniadis, "Social network analysis and social capital in marketing: theory and practical implementation," 2016.

[26] Andreas M. Kaplan, Michael Haenlein, "Users of the world, unite! The challenges and opportunities of Social Media," in *Business Horizons*, 2010.

[27] A. Tugend, "Too many choices: A problem that can paralyze," *The New York Times,* p. 26, 2010.

[28] Jingfeng Cui, Zhaoxia Wang, Seng-Beng Ho, Erik Cambria, "Survey on sentiment analysis: evolution of research methods and topics".

[29] Nhan Cach Dang, María N. Moreno-García, Fernando De la Prieta, "Sentiment Analysis Based on Deep Learning:A Comparative Study," in *MDPI*, 2020.

[30] Socher, R., Perelygin, A., Wu, J. Y., Chuang, J., Manning, C. D., Ng, A. Y., & Potts, C., "Recursive deep models for semantic compositionality over a sentiment treebank. In Conference on Empirical Methods in Natural Language Processing.," 2013.

[31] B. Liu, Sentiment analysis and opinion mining. Synthesis Lectures on Human Language Technologies.

[32] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K., "Pre-training of deep bidirectional transformers for language understanding," in *BERT*.

[33] Gephi, "Docs Gephi," Gephi Org, [Online]. Available: https://docs.gephi.org/.

[34] Bastian, M., Heymann, S., & Jacomy, M., "Gephi: An open source software for exploring and manipulating networks," in *International AAAI Conference on Weblogs and Social Media*, 2009.

[35] "Vedar Sentiment," [Online]. Available: https://vadersentiment.readthedocs.io/en/latest/.

[36] C.J. Hutto, Eric Gilbert, "VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text," in *Eighth International AAAI Conference on Weblogs and Social Media*.

[37] Zhang, Z., Chen, Q., & Liu, B., "Deep Learning for Sentiment Analysis," in *Data Mining and Knowledge Discovery*, 2018.

[38] "Sentiment Analysis: AFINN vs Bert AI Algorithms (using the Twitter and Amazon examples)," [Online]. Available:

https://noduslabs.com/featured/sentiment-analysis-afinn-bert-ai-twitter-amazon/.

[39] Z. Zhang, "Text Mining for Social and Behavioral Research," 2018.

[40] Shefali Singh, Tureen Chauhan, Priyanka Meel, "Mining Tourists' Opinions on Popular Indian Tourism Hotspots using Sentiment Analysis and Topic Modeling," 2021.

7 APPENDICES

Appendix

The works of thesis project is included in following url:
https://github.com/ashanoulu/finnish_recipe_recommender