

UNIVERSITY OF BERGEN

Department of Information Science and Media Studies

MASTERS THESIS

**Novel Methods Using Human Emotion and
Visual Features for Recommending Movies**

Author: Øyvind Johannessen

Supervisor: Mehdi Elahi

Co-supervisor: Marko Tkalčič

June 1, 2021

Abstract

This master thesis investigates novel methods using human emotion as contextual information to estimate and elicit ratings when watching movie trailers. The aim is to acquire user preferences without the intrusive and time-consuming behavior of Explicit Feedback strategies, and generate quality recommendations. The proposed preference-elicitation technique is implemented as an Emotion-based Filtering technique (EF) to generate recommendations, and is evaluated against two other recommendation techniques. One Visual-based Filtering technique, using low-level visual features of movies, and one Collaborative Filtering (CF) using explicit ratings. In terms of *Accuracy*, we found the Emotion-based Filtering technique (EF) to perform better than the two other filtering techniques. In terms of *Diversity*, the Visual-based Filtering (VF) performed best. We further analyse the obtained data to see if movie genres tend to induce specific emotions, and the potential correlation between emotional responses of users and visual features of movie trailers. When investigating emotional responses, we found that *joy* and *disgust* tend to be more prominent in movie genres than other emotions. Our findings also suggest potential correlations on a per movie level. The proposed Emotion-based Filtering technique can be adopted as an Implicit Feedback strategy to obtain user preferences. For future work, we will extend the experiment with more participants and build stronger affective profiles to be studied when recommending movies.

Contents

Abstract	ii
1 Introduction	2
1.1 Background	2
Recommendation Techniques	2
Cold Start Limitation	3
Preference Elicitation	4
1.2 Problem Formulation & Objectives	5
1.3 Preliminary Work	6
1.4 Related work	8
1.5 Approach	10
1.6 Summary	10
2 Methodology	12
2.1 Overview	12
2.2 Material	12
2.3 Environment	14
2.4 Implementation	15
2.5 Architecture Decomposition	16
2.6 Recommender Algorithms	17
2.6.1 Singular Value Decomposition (SVD)	18
2.6.2 Cosine Similarity	19
2.7 Known Issues in Implementation	19

2.8	Pre-study	20
2.8.1	Design	21
2.8.2	Facial Expressions & Prediction Model	23
2.8.3	Features	24
2.8.4	Models	25
2.9	Main study	27
2.9.1	Design	27
	Step 1: Demographics	27
	Step 2: Personality Questionnaire	27
	Step 3: Favourite Genre	28
	Step 4: Selecting Movies	29
	Step 5: Watching and Rating	30
	Step 6: Recommendations	31
	Step 7: System Usability Questionnaire	31
2.10	Shortcomings	32
2.11	Advantages	33
3	Results	34
3.1	Users	35
3.2	Emotion & Visuals	36
3.2.1	Procedure	36
3.2.2	Average Emotion	39
3.2.3	Correlation	42
3.3	Recommendation Evaluation	46
3.3.1	Procedure	47
3.3.2	User Evaluation	50
3.3.3	User Evaluation by Personality	54
3.4	Affective Preference Elicitation	57
3.5	The System Usability	58

- 4 Conclusions** **61**
- 4.1 Summary 61
- 4.2 Discussion 62
- 4.3 Future Work 64
- 4.4 Acknowledgement 65

- References** **65**

Chapter 1

Introduction

1.1 Background

In the pool of information and data currently available, it can be demanding to navigate, filter, and choose content relevant to our preference. These challenges have created a need for systems and algorithms that can help us manage information overload and make better choices relevant to our needs. Content providers rely on Recommendation Systems to provide a better user experience and help users discover and enjoy their content. In this research field, various types of approaches have already been proposed. The majority of these approaches rely heavily on user preferences (e.g., ratings) to understand the needs and tastes of the users and generate relevant recommendations. Hence, they require the user to provide a number of preferences to build an initial profile [15]. The most popular techniques are Collaborative Filtering (CF), Content-based Filtering (CBF) and Hybrid approaches.

Recommendation Techniques

A collaborative filtering approach discovers the relationships and similarities among users by observing their interactions with the items. The approach compares user-profiles and predicts the likelihood that a user can be interested in an item and suggest the items with the highest likelihood. These interactions can be in the form of ratings, pairwise like or dislike, different behaviors that indicate similar users or relevant keywords. Typically, collaborative filtering uses the ratings that users provide to items and recommend the items with the highest predicted ratings. The missing ratings for a user are predicted based on the user-item interactions. The items with the highest score are recommended to the user [12], [2].

Content-based filtering is more centered around the content of the items. This approach relies on content features that describe an item and find content-based similarities between

the items, hence suggests items of interest based on their associated features [35]. Systems that utilize this approach have a greater likelihood of recommending new items. This means that they only need to know what items a specific user has interacted with and recommend similar items. A recommendation model that is based on the content of items has a multitude of feature possibilities. For instance, news recommender systems model the words in the news articles as features and suggest articles with features similar to those the user preferred before. When recommending movies, a content-based approach can utilize the title, genres, author, director, subtitles, or tags. Every feature that describes an item can be used in content-based models and are often referred to as *descriptive attributes* [2].

A Hybrid recommendation approach combines collaborative filtering and content-based filtering in order to utilize the benefits of both. When combining these approaches, systems can more precisely generate recommendations that are relevant to users. They often perform reliably in a variety of environments [6].

In this master thesis, we propose an Emotion-based filtering (EF) technique using ratings estimated from observed emotional responses. This technique is compared and evaluated against a content-based approach using low-level visual features of movie trailers, called a Visual-based Filtering (VF), and a baseline technique using Collaborative Filtering (CF). The proposed technique aims to alleviate some of the challenges in previous techniques.

Cold Start Limitation

One of the main challenges of collaborative filtering is data sparsity and cold-start limitations. This can occur both when a new user enters the system (*new user* problem) and when a new item is added to the catalog (*new item* problem). These situations occur when a new user has not provided any rating to, or interacted with, any item. The recommender system is unable to generate a personalized recommendation. This problem then leads to poor recommendations where the most popular movies are repeatedly recommended to users without taking into account the particular tastes of the user [11]. Furthermore, in a new item situation, a new item often lacks enough ratings to be evaluated as a potential candidate for recommendation.

The cold-start problem is a research field actively searching for novel solutions. The aim is to improve the quality of recommendations for the users when not much data is available. Some solutions employ additional information such as social context, user profiles, or descriptive features to filter content on sparse datasets. With the assumption that similar users have similar tastes, these solutions try overcome the cold-start limitations based on contextual information besides the actual preferences of users. While such techniques can limit

the cold-start, they depend on finding similarities that are not based on preferences related to the system content, hence inherit limitations [20, 3]. While there are many solutions for the problem, all of them inherit limitations, and none can completely resolve the problem. Hence, there is a need for new research to address the limitations better and propose more effective solutions [29], [37], [26], [13].

Preference Elicitation

Preference elicitation is a category of strategies to acquire user preferences. The root of these strategies originates from Machine Learning techniques, specifically *Active Learning*. Eliciting preferences is often performed using two approaches. The first approach is called Explicit Feedback, while the second approach is called Implicit Feedback. When using explicit feedback, the recommender system require the user to provide ratings actively. This approach often yields good results but require efforts from the user. This is often perceived as intrusive and time-consuming. Instead of actively asking users to provide preferences, implicit feedback infers user preferences by monitoring their actions within the system. This approach is reported to be less accurate than explicit feedback, but does not intrude or require users to actively provide preferences [24].

Preference elicitation is either applied at an early stage when a user registers, or later when the user has already started interacting with the system [36]. One example when obtaining preferences is to provide a new user with a list of items to rate. This is often applied when the user registers. As an example, in Netflix¹, which is a popular movie streaming service that provides its subscribers with up to one million movies, active learning is applied in the beginning to obtain initial data from users where the system asks the users to choose and rate a few movies before receiving any personalized recommendation. Another example is using a conversational approach where the system tries to motivate the user to rate more freely without asking explicitly. In contrast to the conversational approach, Decision-Tree-Based methods try to identify items that provide knowledge about the prediction error within recommendation systems. These items are actively requested to be rated to increase accuracy [17], [34]. Some approaches are more contextual and try to utilize user characteristics (e.g., user personality) to predict and obtain a better quality of ratings [16], [18].

¹<https://www.netflix.com/no/>

1.2 Problem Formulation & Objectives

In the context of cold-start and data sparsity, preference elicitation strategies are valuable to overcome these limitations, and generate personalized recommendations. The intrusive and time-consuming aspect of explicit feedback reduces the engagement of providing preferences, and leads to data sparsity and poor recommendations. On the other hand, implicit feedback has proven to be less accurate in obtaining quality preferences.

In addressing this problem, this thesis proposes a novel methodology that can be adopted in order to obtain user preferences (e.g., ratings). The main objective is to collect quality preference, without requesting item ratings. Promising new fields of research are looking at Affective Computing and how emotional responses captured from facial expressions can contribute in generating good recommendations. These emotional responses can be used as contextual data through monitoring the user, hence obtain implicit feedback to infer preferences. Affective Computing libraries give us the opportunity to detect and collect facial expressions and emotions captured in each frame of web cameras. By collecting and processing these data, this research propose an implicit feedback strategy to estimate and elicit ratings continuously without explicitly asking users to rate items. Hence, the proposed approach may substantially alleviate data sparsity and the intrusive aspect of preference elicitation.

In addition, this thesis represents a feasibility study based on a new line of research that aims at presenting findings and relationships between visual features of movies, and emotional responses of users. The discoveries made can contribute to further research in generating recommendations. The use of visual features and emotions to generate recommendations are the components used in this experiment to call it an *Emotion-based Movie Recommender*.

In order to achieve the above noted objective, a number of objectives have been formulated:

1. Developing the prototype of the Emotion-based Movie Recommender, a novel approach eliciting user preferences (i.e., ratings) without requiring item ratings, as well as generating quality recommendations based on these preferences.
2. Building a Machine Learning model that can predict the preferences of users based on their emotional responses extracted from their facial expressions.
3. Evaluating the Emotion-based Filtering (EF) approach and comparing it with other techniques, i.e., Collaborative Filtering (CF), and Visual-based Filtering (VF).
4. Investigating the potential correlation between the emotional response of the users and visual features encoded within frame-by-frames of the movies.

From the objectives noted above, a number of research questions have been formulated:

- **RQ1:** Are there any difference among movie genres in terms of the emotional responses obtained from facial expressions of the users?
- **RQ2:** Is there a correlation between visual features encapsulated within movies and the emotional responses the users express?
- **RQ3:** In terms of Accuracy and Diversity, what is the quality of recommendation based on emotional responses, using facial expressions, in comparison to the other approaches?
- **RQ4:** In terms of Accuracy and Diversity, do users with similar personality traits prefer similar recommendation approaches?
- **RQ5:** Can the preferences of users be elicited from their emotional responses extracted from the facial expressions in order to generate movie recommendations?

1.3 Preliminary Work

The section presents preliminary work which has been used as directional guidelines for this research. There have been attempts to incorporate emotions into recommender systems with the goal of improving recommendation generation. These recommender systems are referred to as Affective Recommender Systems. Tkalčič et al. [42] has surveyed several research attempts in developing Affective Recommender Systems. They found that most of the research has been conducted independently and stretches across recommender systems and affective computing. Tkalčič et al. [42] presents a unifying framework to help researchers position their work when researching Affective Recommender Systems. The framework aims at unifying research conducted in the community so that researchers can benefit from each others work.

Within the unifying framework, Tkalčič et al. [42] presents three stages that cover the detection of a user's emotional state and how it might be used in the context of Affective Recommender Systems. The stages are listed as *the entry stage*, *the consumption stage*, and *the exit stage*. The stages organizes the affective state of users into categories, which can help to understand how and when it fits into a recommender systems. *The entry stage* revolves around understanding a users mood when they first enter the system. This stage carries with it a mood that has been caused by something outside the context of the recommender system, and can be utilized to recommend movies. When a user has selected content to consume based on the entry mood, *the consumption stage* is entered, which is when the content induce affective responses in the user. In the context of watching movie trailers, the emotions

captured in the consumption stage are continuous and change over time. The last stage in the unifying framework is the *exit stage*. This stage revolves around users mood after they have consumed the content. The mood will influence the users next action, and if the user continues to use the recommender system, the mood captured in the *exit stage* can be used as the new entry mood for *the entry stage*. Tkalcic et al. [42] explains that the detection of the exit mood can be used as an unobtrusive technique to collect feedback. With the knowledge obtained in the survey, Tkalcic et al. [42] presents important research areas to work on.

Tkalcic et al. [42] identified four areas for further research to develop Affective Recommender Systems. They recommend researching the use of emotions as context in the *entry stage*, modeling affective content profiles, recommending content using affective profiles, and building a set of datasets. The four areas are noted as important to drive the research of affective recommender systems forward.

When conducting research in the context of Affective Recommender Systems, it is important to make good use of the emotional states. In order to better understand emotional states, two models have been proposed. The first model is called *The Universal Emotions Model*, and the second model is called *The Dimensional Model*. Within the universal emotions model, users emotional states are categorized as a set of universal emotions. The most popular names for these categories are happiness, anger, sadness, fear, disgust, and surprise. In contrast to the universal emotions model, the dimensional model uses a multidimensional space to describe the quality of emotion. These dimensions are often named valence, arousal, and dominance (VAD), where valence describes the pleasantness of a stimulus, arousal describes the intensity of emotion, and dominance describes the level of control over the emotional state. These models map emotions to categories and help interpretation of the emotional state [42], [43].

In the context of movie trailers, the consumption stage consists of watching trailers which possibly induce emotions affected by visual features. Moghaddam et al. [32] explains that visual features are "low-level" features. These are visual features in the movie and not any external features like genre or year. Every movie is split into shots. These shots contain frames. One frame is selected from each shot as a representative called a Key-frame. These key-frames contain visual features that are extracted and analyzed. The features extracted are aggregated into a vector that represents the visual feature of a movie. The visual features characterize a movies attractiveness and are explained by Moghaddam et al. [32]:

- **Sharpness** measures the clarity and level of details within the elements of a frame.
- **Sharpness Variation** is calculated via the standard deviation of all pixel sharpness values.

- **Contrast** measures the relative difference in brightness or color of local features in a frame.
- **RGB Contrast** is contrast, but extended to a three-dimensional RGB color space.
- **Saturation** measures the colorfulness of the frame relative to the brightness.
- **Saturation Variation** measures the variation in saturation via the sample standard deviation of all pixel saturation in a frame.
- **Brightness** measures the average brightness of a frame.
- **Colorfulness** measures the individual color distance of the pixels in a frame.
- **Entropy** determine how much information needs to be encoded by a compression algorithm.
- **Naturalness** measures the difference (or similarity) between a frame and the human visual perception of the real world, with respect to colorfulness and dynamic range.

1.4 Related work

This section will cover some of the related work exploring the use of emotional responses and visual features. The use of affective data has given rise to interesting research and promising results within the context of multimedia content. Joho et al. [25] used facial expressions to detect affective highlights of videos to which they computed personalized affective summarisations. The need for affective summarisation of multimedia content is helpful for rapid comprehension of video content. Affective summaries also assist users in deciding to watch a whole video or not. They used facial expressions and content-based features in two separate models to generate video summaries. They compared the performance of the models against the viewer's annotation of highlights. Joho et al. [25] found that using facial expressions to generate personalized movie summaries was challenging and did not perform to the level of expected satisfaction. On the other hand, in contrast to content-based techniques, Joho et al. [25] concludes that using facial expressions has the potential to generate personalized summaries of video content.

With focusing on music content rather than video content, Tkalčič et al. [40] conducted research in detecting emotional responses through facial expressions when listening to music. In the study, they used a pairwise approach where the user could decide which of two songs they liked the most. While the user was listening, they captured the facial expressions of the user. The authors found that it is possible to predict the preferred song by interpreting their

emotions. They also found the approach to be better in prediction than the more common method of calculating the time user spends listening to a song.

From the findings presented by Tkalčič et al. [40], it is natural to assume that both sound waves and visual features can impact our affective state. To analyse this assumption with the focus on video content, we can extract the visual features from movies and further analyse their potential applications.

By analysing these features, promising results have been found. In the paper Predicting Movie Popularity and Ratings with Visual Features, Moghaddam et al. [32] presents a more accurate method for predicting movie popularity based on visual features. Moghaddam et al. [32] used 13'000 movie trailers and managed to demonstrate promising results. The research shows alternative ways of handling the cold-start limitation, specifically *new item* problem. The use of visual features or the *attractiveness* of a movie can help newer movies get recommended. Visual features in movies have a correlation with average ratings and number of ratings, and can be used to predict a movie's estimated popularity score. With this knowledge, Moghaddam et al. [32] explains that it might be possible to predict the popularity of a film before its release.

While visual features have been proven applicable in predicting movie popularity and ratings, they have also been utilized in recommender systems as descriptive content features with encouraging outcomes. Deldjoo et al. [8] experimented with using visual features to generate recommendations. They used two approaches to extract visual features. First, they extracted visual features from MPEG-7 descriptors, and secondly extracted visual features automatically using deep learning. The features collected from the two approaches were used in the recommender algorithm to generate recommendations. In comparison with using human-generated features such as genre and tags, Deldjoo et al. [8] found that using automatically extracted visual features can generate equally good recommendations. Deldjoo et al. [8] explains how automatically extracted visual features introduce flexibility when handling new items which humans have not yet tagged. Furthermore, Deldjoo et al. [8] found that recommendations were consistently better than the baseline approaches when using the visual features extracted from MPEG-7 descriptors.

In a similar experiment previously conducted, Deldjoo et al. [7] reports that the automatically extracted low-level visual features achieved better results than high-level features such as movie genre. They also found that extracting visual features from movie trailers instead of full-length movies produced better results than the baselines. These findings make way for researching opportunities when constrained to short-length videos.

In summary, the use of Affective Computing and mapping of emotional responses has given rise to intriguing research fields which can contribute in developing Affective Recommender

Systems. The use of visual features has been proven to be a useful source to generate low-level features which generate quality recommendation without the need to rely on high-level, human-annotated features. The assumption that multimedia content induce a continual change in emotional responses brings intriguing questions on the relationship between visual features and emotional responses, and if they can be utilized to generate quality recommendations.

1.5 Approach

To achieve the previously mentioned objectives, the approach in this research relies on established technologies and evaluation methods. While detailed descriptions of materials and procedures are presented in the methodology chapter, this section provides an overview of the approach. To develop the Emotion-based Movie Recommender, this research relies on Design Science Methodology guidelines [23]. This methodology presents a framework and guidelines to evaluate research when developing Information Systems. Inspired by these guidelines, a new innovative artifact is built and instantiated which acts as a provisional preference elicitation application. The artifact is developed rigorously similar to established methods previously tested. The experiment-design and evaluation-surveys are adopted from [9] and guided by Nielsen's heuristics [33]. In contrast to the recommendation techniques evaluated in [9], this research evaluates three different recommendation techniques. The first technique is the proposed Emotion-based Filtering (EF), which rely on ratings estimated through observed facial expressions and emotions, hence implicit feedback. The second technique is Visual-based Filtering (VF) to find similar movies based on low-level visual features of movie trailers. The third technique is Collaborative Filtering (CF) using ratings obtained through explicit feedback. These three recommendation approaches are evaluated by users considering the perceived Accuracy and Diversity. The integration of Affective Computing to capture facial expressions and emotions are adopted from [40]. In addition, the processing of these data, and the prediction of ratings rigorously rely on established programming libraries.

1.6 Summary

This chapter has presented the background, problem formulation, objectives, preliminary work, and related work. The first part addressed the importance of techniques that help users filter and make decisions when selecting content. With this, short explanations of the most common approaches such as Collaborative Filtering (CF), Content-based Filter-

ing (CBF), and Hybrid approaches were explained together with some of the challenges they incorporate.

The cold-start limitation is one of those challenges and has several solutions. While no solution has yet to resolve the problem completely, many strategies have been proposed. In this regard, preference elicitation techniques to obtain user preferences were discussed. While there are many techniques in preference elicitation, Explicit Feedback and Implicit Feedback were presented as the two categories which define the type of preferences collected. Implicit feedback is often perceived as unobtrusive and less time-consuming but lacks quality. This led to problem formulation and objectives of this research. This research proposes a novel method using Affective Computing to capture facial expressions and emotions to estimate ratings and generate quality recommendations. This approach eliminates the intrusive aspect of explicit preference elicitation and may substantially alleviate data sparsity. In addition, an explanation on how this research will contribute to a new line of research aimed at finding relationships between visual features and emotions were presented.

In the following chapter, the material used, and procedures conducted in this research is presented. The chapter includes a detailed description of how the artifact and experiment was designed and implemented. The chapter is split into two scopes, where the first scope is the *pre-study*, while the second scope explain the *main-study*. The *pre-study* presents the material and procedures applied when collecting facial expression data to develop a model which predicts ratings. Proceeding from the *pre-study*, the *main-study* presents the material and procedures used when implementing the results obtained from the *pre-study* and the additional components required by the artifact. After the methodology chapter, the results is presented. The result chapter presents the results and findings obtained from the main experiment. The final chapter summarizes the research, discusses the findings, and provides recommendations for future work.

Chapter 2

Methodology

2.1 Overview

This chapter presents the main components and procedures used to build, execute and collect data in the experiment. This chapter is composed of four major parts. The first part presents the material, environment, and architecture used. Secondly, the procedures conducted in the *pre-study* when collecting facial expressions is explained together with how these facial expressions were used to train a prediction model. When the *pre-study* is presented, the chapter transition to the *main-study*. The *main-study* presents the design and procedures of the system which was used to conduct the experiment. The last section in this chapter covers some of the shortcomings and advantages found in developing the system.

- The source code for this system can be found on GitHub at:
<https://github.com/Vlummy/Movie-Recommender-System-Emotion-based-Filtering-EF->
- Datasets and Jupyter Notebooks can be found on Bitbucket at:
https://bitbucket.org/Vlum_/emotion-based-filtering-notebooks/src/main/
- The system is online at:
<https://rsa-pi.herokuapp.com/>

2.2 Material

The materials are split into columns representing the category of material and the concrete name of the material used. The material covers the libraries, services, and datasets used to realize the artifact used in this research.

Table 2.1 and table 2.2 presents the material in the columns *frontend*, *backend*, *services*, *primary data* and *secondary data*.

Frontend	Backend	Services
VueJS Framework	Python's Django Framework	Affectiva Facial Expression API
	PostgreSQL Database	Youtube API
	Surprise Library	The Movie Database (TMDb) API
	Sklearn Library	

Table 2.1: Frameworks and services used to build the artifact.

Primary Data	Secondary Data
Facial Expressions	MovieLens 1M Dataset (ml-1m) [22]
	Visual Features (MA14KD_[ORIGINAL]) [14]
	Visual Features (MA14KD_[AGGREGATED]) [14]

Table 2.2: **Primary Data:** Collected in this research. **Secondary Data:** Previously collected Data by external researchers.

Affectiva¹ was used in order to develop the Emotion Based Recommender System. Affectiva is an API (Application Programming Interface) that provides affective computing of facial expressions, emotions, and dimensions of stimulus such as valence, arousal and dominance (VAD). The implementation of Affectiva is based on the guidelines presented by Tkalčič et al. [41]. The data detected by the Affectiva API when watching trailers were used to predict ratings, and together with the Visual Features (MA14KD) dataset [32] provide the basis for analyzing the relationship between visual features and expressed emotions. Figure 2.1 presents a high-level flowchart of the communication between the artifact's modules. Affectiva was used as a middleware to communicate with the frontend and capture facial expression data to be processed by the backend to build affective profiles and elicit ratings.

¹<https://www.affectiva.com/>

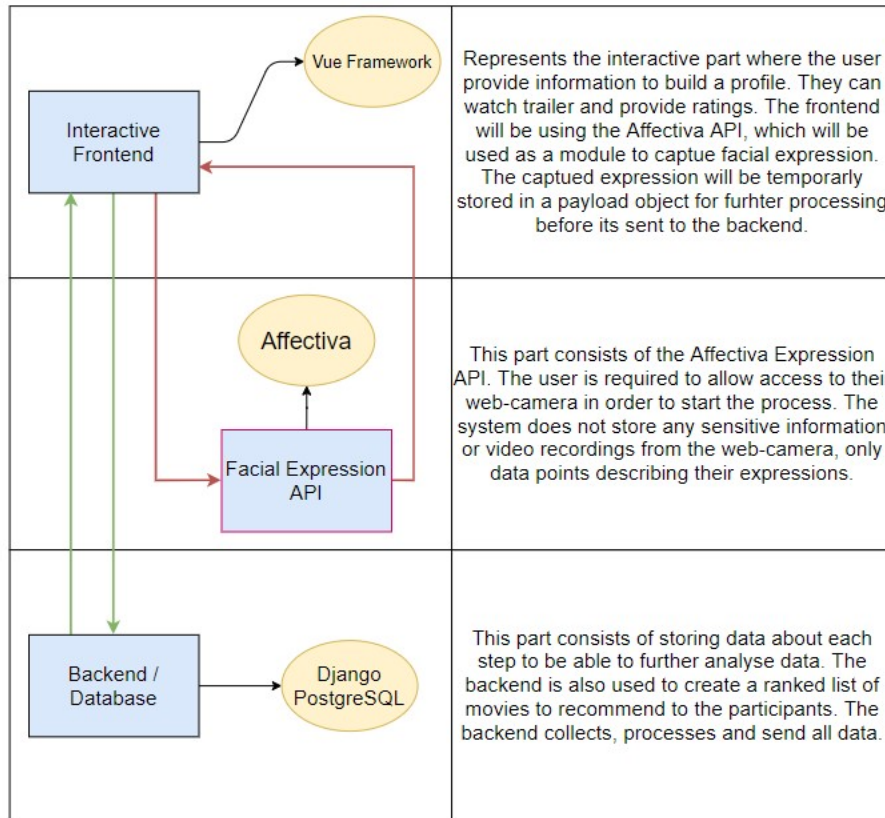


Figure 2.1: High-level model of the artifact components.

2.3 Environment

In order to process all the data, three major components were used. The complete system was built using python's *Django* framework for the backend server, and NodeJS together with VueJS framework as frontend. The backend communicated with a *PostgreSQL* database to store and retrieve data. The VueJS frontend was transpiled to Vanilla Javascript bundle together with a HTML entry point. The backend served the html file as a single page web application. The complete system was deployed as a bundle containing the backend and frontend to a single host machine. The communication between the client and server was implemented using RESTful API endpoints. The frontend and backend communicated using JSON (Javascript Object Notation), which was parsed to dictionary object to be further processed using python backend.

2.4 Implementation

The system was built for the purpose of online use where any user could register, build a user profile and get recommendations. The implementation required a internet connection and a computer. The system was not configured to work well with mobile or tablet devices since we needed clear and precise capture of faces through web camera. In order to interact with the system, the user needed a computer with a web browser and a web camera.

The backend was implemented as a series of decoupled services with encapsulated responsibilities to handle *user services*, *movie services*, *data services*, *affective services*, and *recommendation services*. The *user service* processed and provided functions in relation to access tokens, profile data, passwords and additional user related data. The *data service* handled third party APIs used to search and retrieve movie trailers from Google's Youtube API. The *movie service* contained several functions for retrieving and filtering movies from the ratings dataset, visual features dataset, and additional information such as movie titles and movie genres. The *affective service* handled the storing of sessions containing facial expressions of users, and the method used to predict ratings and store the predicted rating to a users rating profile. The *recommendation service* contained functions to generate, filter and provide lists of movie recommendations. Each service module was accessed through a series of API endpoints where each service had a root resource function that operated based on the parameters provided. The next section presents the architecture decomposition with focus on usage flow in the system.

2.5 Architecture Decomposition

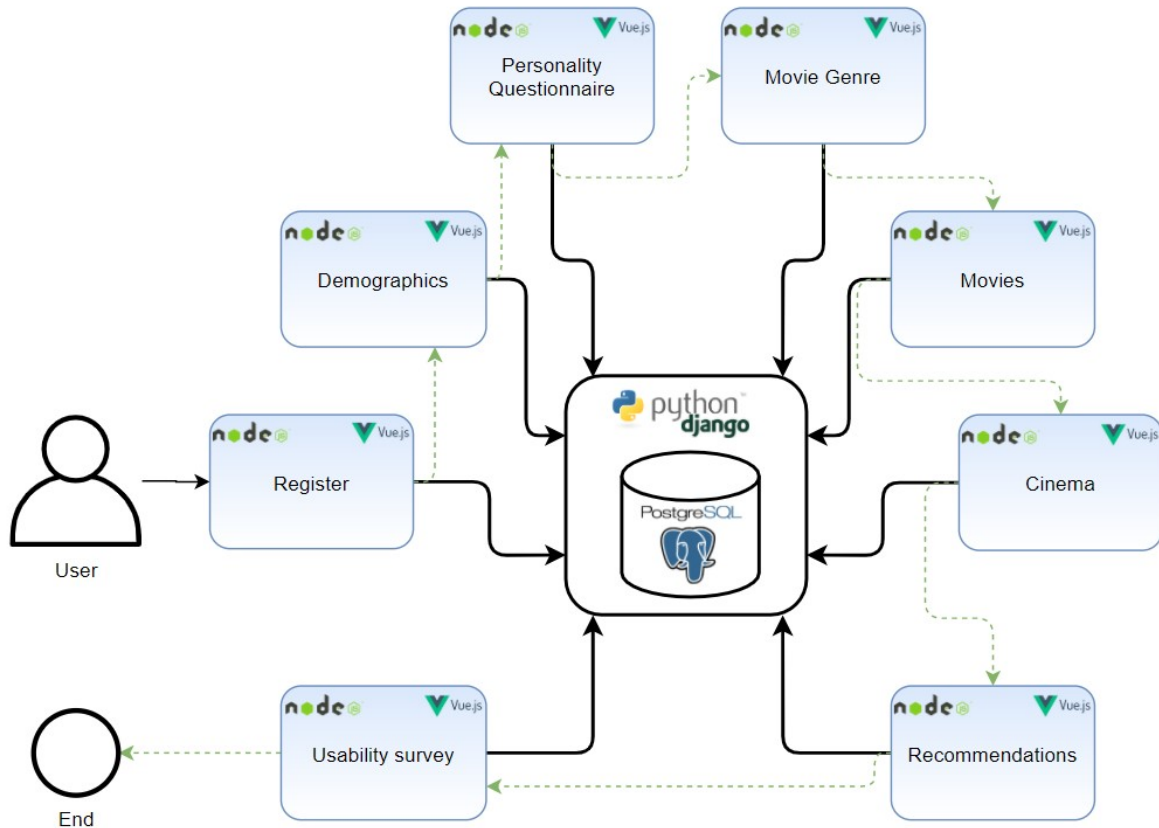
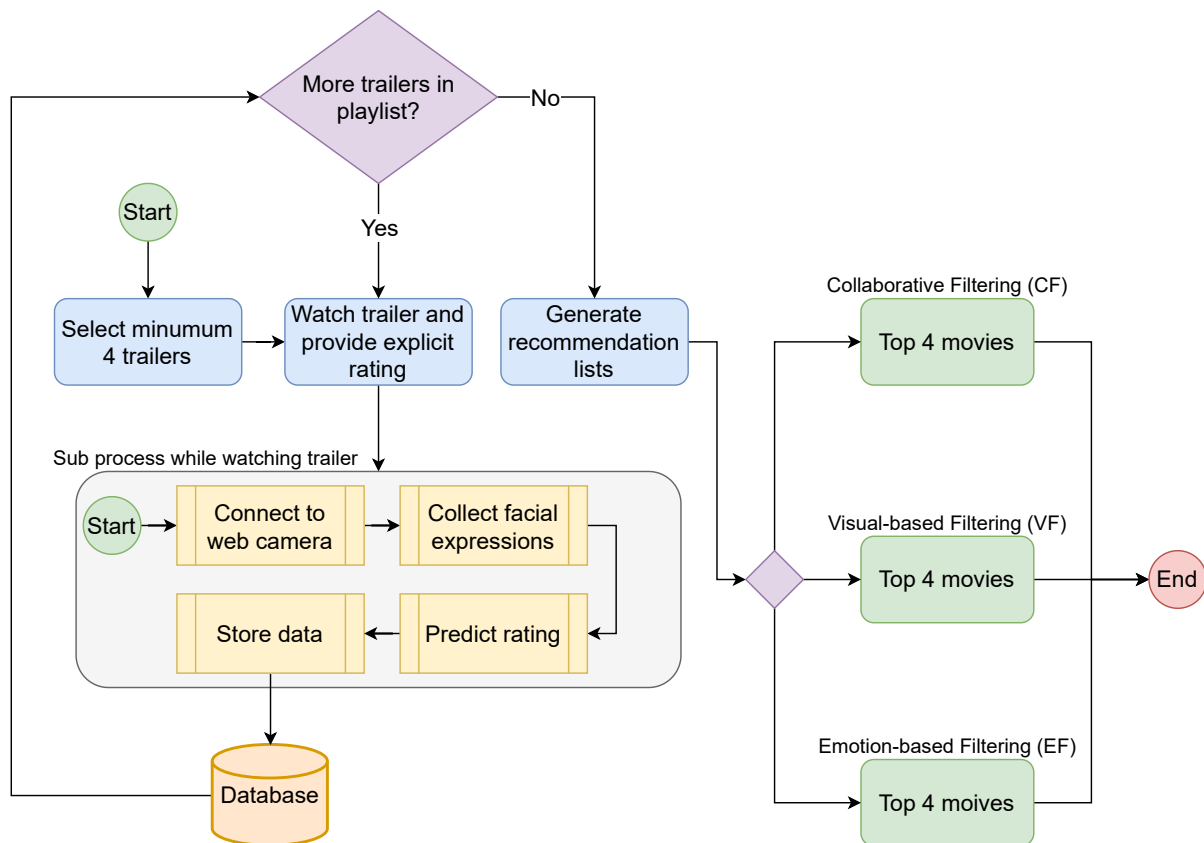


Figure 2.2: Decomposition model: Demonstrates each step of the experiment when interacting with the Emotion-based Movie Recommender artifact.

Figure 2.2 demonstrates the usage flow and each step provided to the user when interacting with the system. The system is built as a web application where the user interacts with the system through a web browser. The design is an adaptation of the work conducted by Deldjoo et al. [9]. The frontend is split into sequential steps. The user needs to complete the instructions on each step in order to continue to the next step. When a step has been completed, the information provided by the user is passed as a payload to the backend for further processing and storing. The *Cinema* step is where the user watches and rate trailers. The step that follows the *Cinema* step consists of receiving and evaluating three recommendation lists based on the algorithms: Singular Value Decomposition (SVD) for Emotion-based Filtering (EF) and Collaborative Filtering (CF), and Cosine Similarity Measure for Visual-based Filtering (VF).

The process of capturing facial expressions, predicting a rating and using the data to recommend movies is explained in figure 2.3. The flowchart demonstrates the architecture from selecting movie trailers, sub-procedures that are executed when watching trailers, and the

process of generating recommendations. When the user is watching a movie trailers, the facial expressions and emotions are captured, processed, and used to predict a rating which is further stored in the users rating profile. After collecting ratings through explicit feedback, predicting ratings from facial expressions and emotions, and collecting visual features for the watched movie trailers, the system uses the obtained preferences to generate recommendations.



(a) Flow of the recommendation architecture

Figure 2.3: Preference Elicitation for the Emotion-based Filtering technique: From selecting movies, watching trailers and receive recommendations. The sub process demonstrates how facial expressions are captured and used to predict ratings, and build affective user profiles.

2.6 Recommender Algorithms

Established libraries were used to generate recommendations from the obtained ratings and the visual features. For the Emotion-based Filtering (EF) and Collaborative Filtering (CF) approach, we used Surprise library². Surprise library contains an implementation of the Singular Value Decomposition (SVD) algorithm, which was trained using 100 000 movie ratings randomly sampled from the MovieLens 1M Dataset (ml-1m) [22]. The algorithm was trained

²<http://surpriselib.com/>

using GridSearchCV, which finds the model with the best parameters and provides an accuracy based on cross-validation using five folds. Table 2.3 presents the score from training the SVD model.

Model	RMSE Score	MAE Score	Parameters	Parameter Value	Parameter Description
SVD	0.958	0.768	n_epochs	10	The number of iterations of the SGD procedure.
			lr_all	0.005	The learning rate for all parameters.
			reg_all	0.04	The regularization term for all parameters.

Table 2.3: Parameters and score from training the SVD algorithm. RMSE: Root Mean Squared Error. MAE: Mean Absolute Error.

For the Visual-based Filtering technique (VF), Scikit-learn library³ was used. From the metrics module in Sklearn library, the cosine similarity⁴ function was applied to calculate a similarity matrix between each movie the user had seen and all the movies in the dataset. The similarity was based on the visual features of the input movie and all other movies. From the generated list of similarity scores, the most similar movie was selected as a recommendation candidate. Table 2.4 depicts an example where *The Return of Ringo (1965)* is considered the most similar movie to *A Modern Affair (1995)* in terms of visual features.

index	Similarity	MovieId	MovieTitle	Selected Title
360	1.000000	623	Modern Affair, A (1995)	Modern Affair, A (1995)
12734	0.999876	118752	The Return of Ringo (1965)	Modern Affair, A (1995)
2030	0.999859	3144	Glass Bottom Boat, The (1966)	Modern Affair, A (1995)
1139	0.999854	1878	Woo (1998)	Modern Affair, A (1995)
13689	0.999844	128369	The Anderssons in Greece: All Inclusive (2012)	Modern Affair, A (1995)

Table 2.4: Top 4 similar movies to the movie Modern Affair, A (1995)

To further provide some insights to how these algorithms operate, a brief explanation is presented in the next section.

2.6.1 Singular Value Decomposition (SVD)

Singular Value Decomposition (SVD) is a popular algorithm used for Collaborative Filtering (CF). SVD is a matrix factorization algorithm that finds the latent feature values of users and

³<https://scikit-learn.org/>

⁴https://scikit-learn.org/stable/modules/generated/sklearn.metrics.pairwise.cosine_similarity.html

items. The algorithm predicts unknown scores by using the known features, which are inferred by user ratings. Two matrices are found by the algorithm. One matrix represents the score between *items* and *features*. The second matrix is the score between *users* and *features*. The score between them are represented as real numbers. The two matrices can be formulated as each having a *f-dimension feature vector*:

$$I = \text{Item}$$

$$U = \text{User}$$

$$f = \text{feature}$$

$$I \in \mathbb{R}^{f \times n}$$

$$U \in \mathbb{R}^{f \times n}$$

The values V between users and items are estimated by using a prediction function p on the *user-features* vector and the *item-features* vector.

$$V = \text{Value}$$

$$u = \text{user}$$

$$i = \text{item}$$

$$V \in \mathbb{R}^{u \times i}$$

$$V_{ij} = p(U_i, M_j)$$

The algorithm is optimized by minimizing the sum of squared errors between existing scores and their prediction values [31], [44].

2.6.2 Cosine Similarity

Cosine similarity is a method to calculate how similar objects are using a set of features to compare. Cosine similarity computes the normalized dot-product between two vectors, where the normalization is Euclidean (L2-norm). The similarity score is presented within the range of 0 and 1 between each sample. The formula is denoted as:

$$\text{cosine}(a, b) = \frac{a \times b}{\|a\| \times \|b\|}$$

In the formula above, the numerator calculates the dot-product, and is regularized by the L2-norm for a and b [38].

2.7 Known Issues in Implementation

The implementation of the different components brought with it some challenges along the way. These challenges required both logical and technical solutions. One challenge revolved

around letting the user feel, as much as possible, free to navigate the system, and at the same time, be notified when valuable data was about to be unreliable. If a user watched a trailer with a faulty camera or went away from the computer during a trailer, the system was prone to fail because of the missing facial expression. The system was highly dependent that the user stayed focused and undisturbed through the whole experiment. To ensure that the system detected facial expressions and did not continue trying to predict a rating on little to no data, an interval clock was implemented that checked for an increase in data each 20th second. This means that if there is an increase in incoming facial expressions, the system will continue to play the trailer and collect facial expressions; else, it pauses the trailer and the facial expression detection while notifying the user that there is no face to detect.

In addition to handling situations of unreliable facial expressions data, a user had to select and watch minimum four trailers in order to generate reliable recommendations. At best, one user chose more movie trailers than four, but for the system to generate a list of movies, at least four was required. The Visual-based Filtering (VF) model created a recommendation based on each of the four trailers, or the four highest-rated if they watched more. If a user only watched three movies, then the recommendation lists would not match in length with the recommendation lists generated by the two other techniques, and it would be hard for a user to justify one list over another.

The system relied on several APIs to retrieve movie posters and trailers. With these APIs, some challenges did occur. When using Google's YouTube API, there was a slight chance that the trailer retrieved was not the correct trailer. With adjusting the parameters of the request URL to the YouTube API, this problem rarely occurred. The Movie Database (TMDb)⁵ API was used to retrieve metadata such as movie plots and poster images. In some movies, the API did not contain the *TMBb id* which was in the dataset. The missing data resulted in cases where some rare and old movies did not have any posters available. The biggest issue with missing posters was when they got included in the recommendation lists. The user could only see the movie titles and not the movie posters and plot. While measures were taken to reduce these issues, they were never eliminated.

2.8 Pre-study

This section presents the procedures taken to develop a model which predict ratings. The first part presents the user flow and how the facial expressions were collected, and the second part presents the procedures in developing the prediction model and the results obtained.

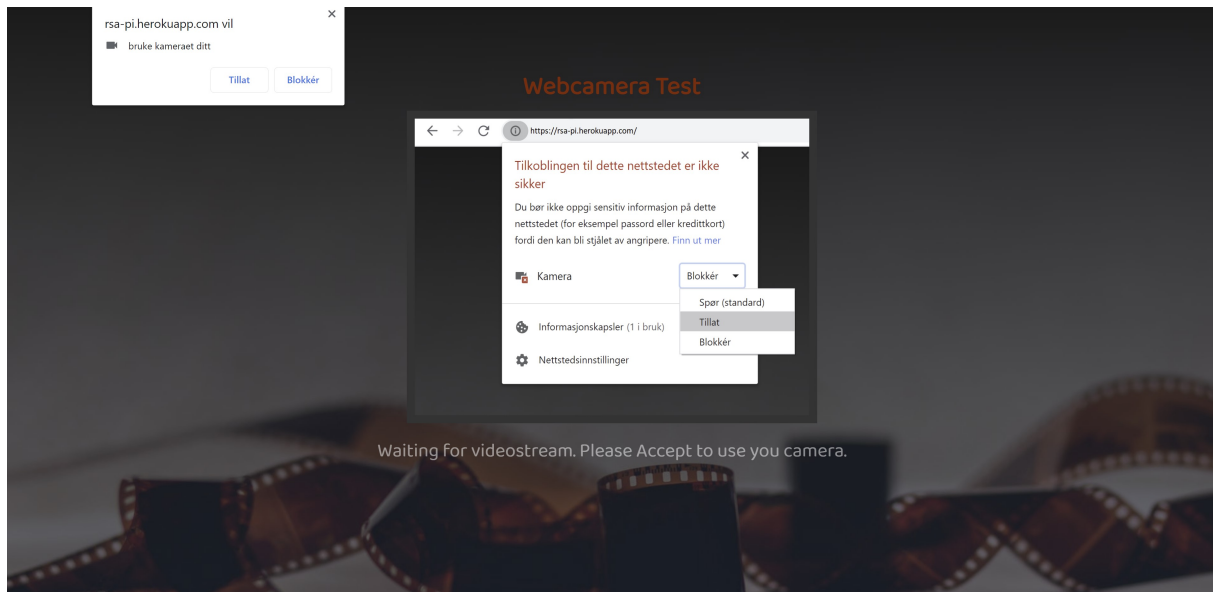
The lack of facial expressions and a prediction model to be used in the *main-study*, drove the

⁵<https://www.themoviedb.org/>

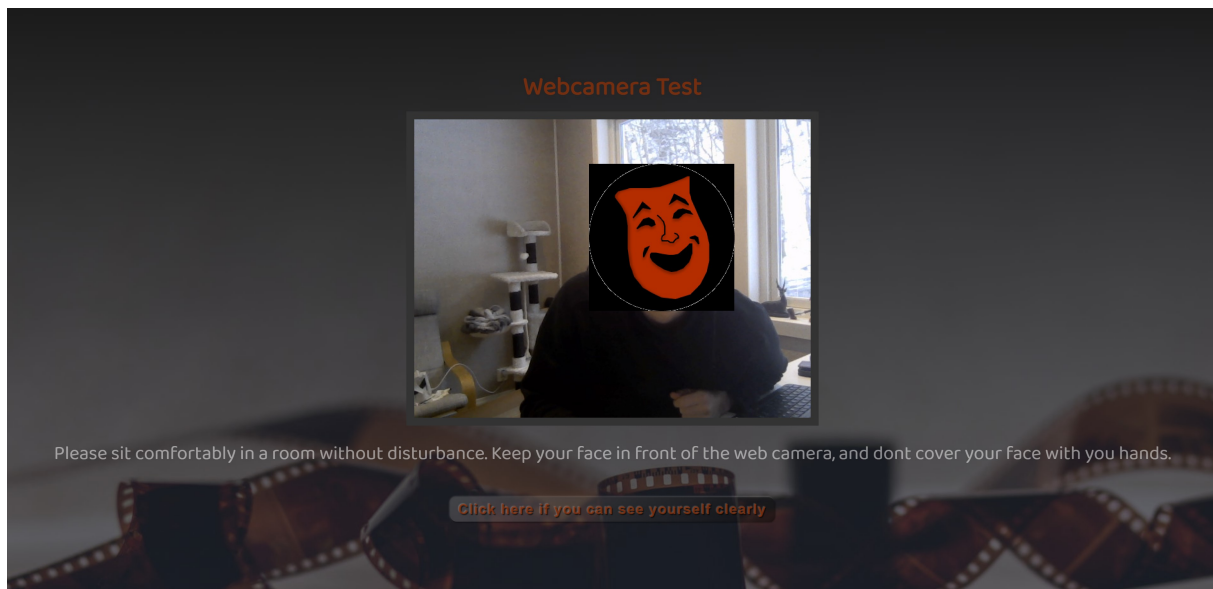
decision to conduct the *pre-study*. In contrast to the *main-study*, the goal of the *pre-study* was reducing the user flow to make it easy for users to select, watch, and rate movies.

2.8.1 Design

Step 1, depicted in the images 2.4, presents the first window a user meets when entering the system. This window helps the user in providing and verifying access to the web camera.



(a) Before allowing access

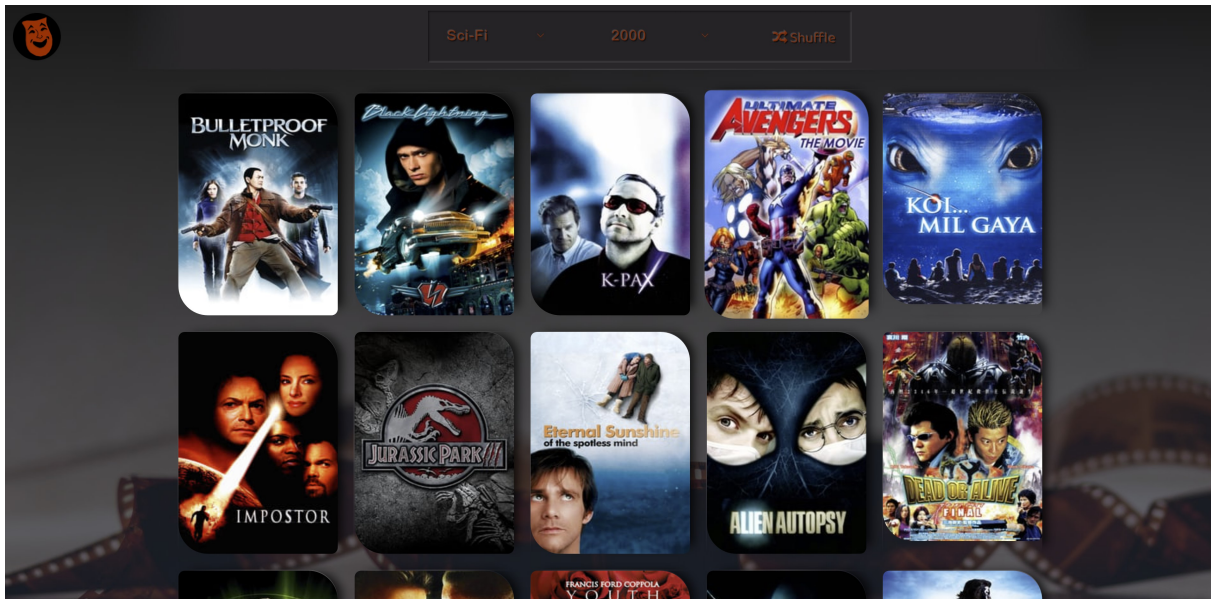


(b) After allowing access

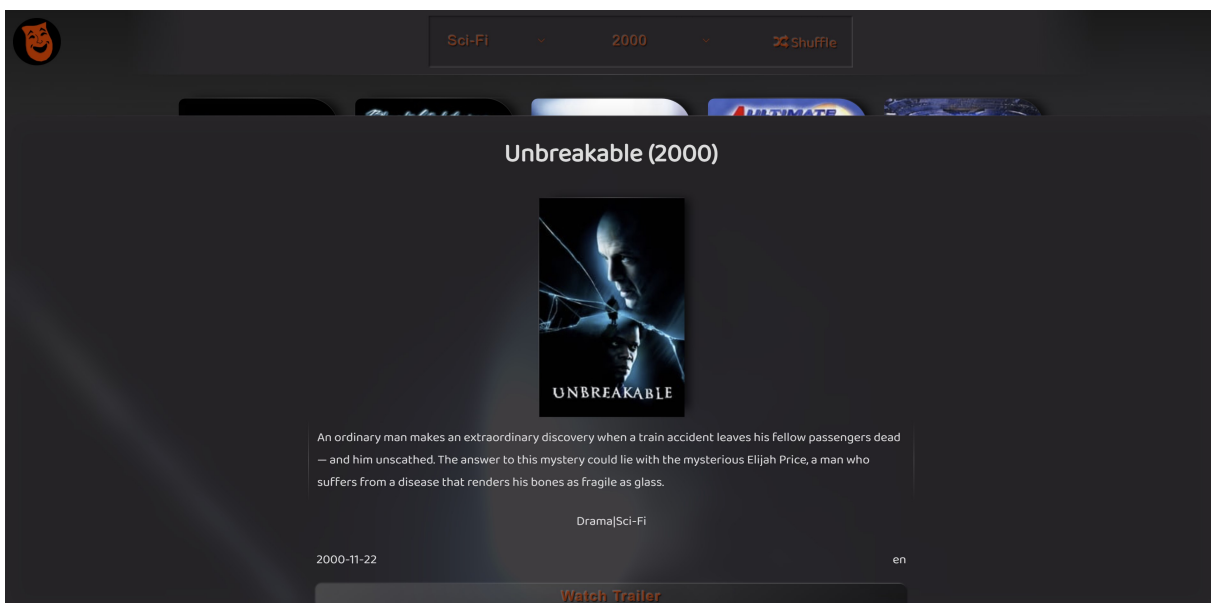
Figure 2.4: Figures of initial landing page

Step 2 depicted in the images 2.5, presents the main dashboard. The dashboard is used to

filter, select, and read about movie plots. Each of the movies has a button which start the movie trailer and the facial expression detection.



(a) Movie grid



(b) Selected movie content

Figure 2.5: Figures of the system in phase 1

Step 3, depicted in the image 2.6, presents the window called the Cinema View. The cinema plays the selected movie trailer, collects facial expressions and receive explicit ratings provided by users.

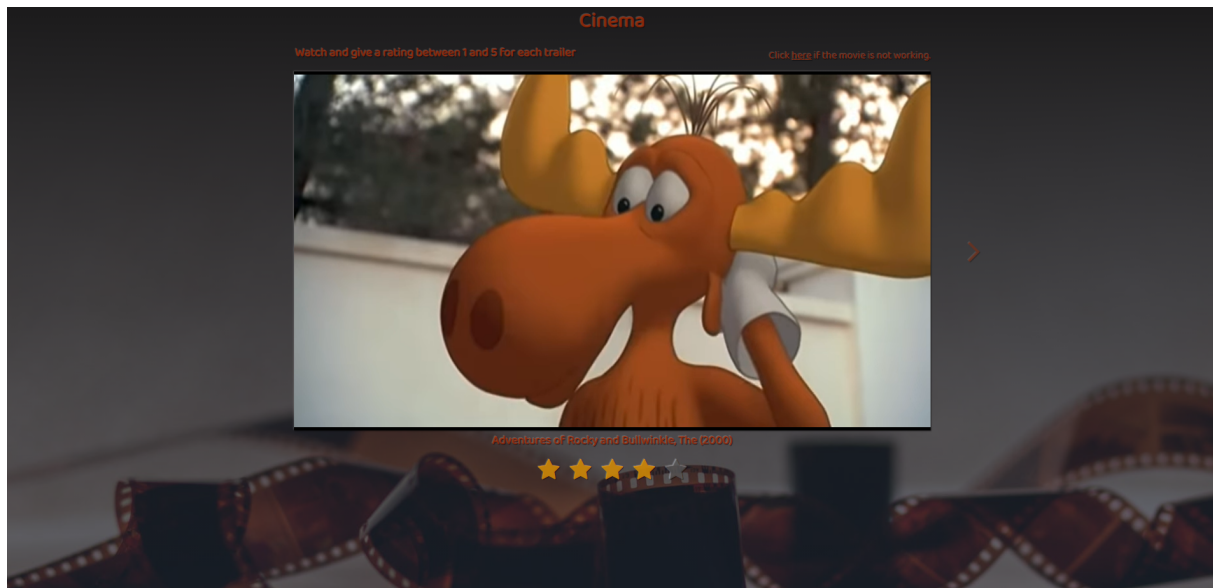
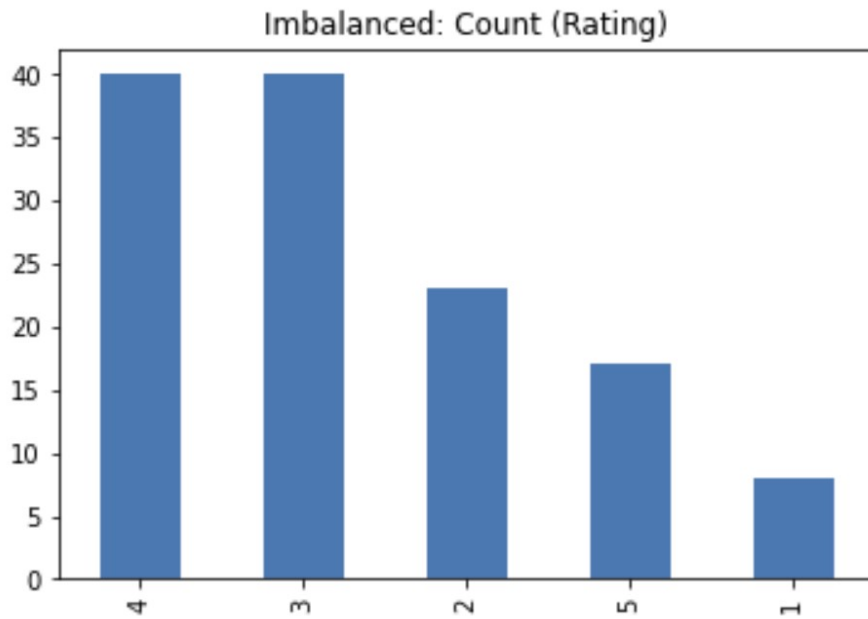


Figure 2.6: Figure of the cinema

2.8.2 Facial Expressions & Prediction Model

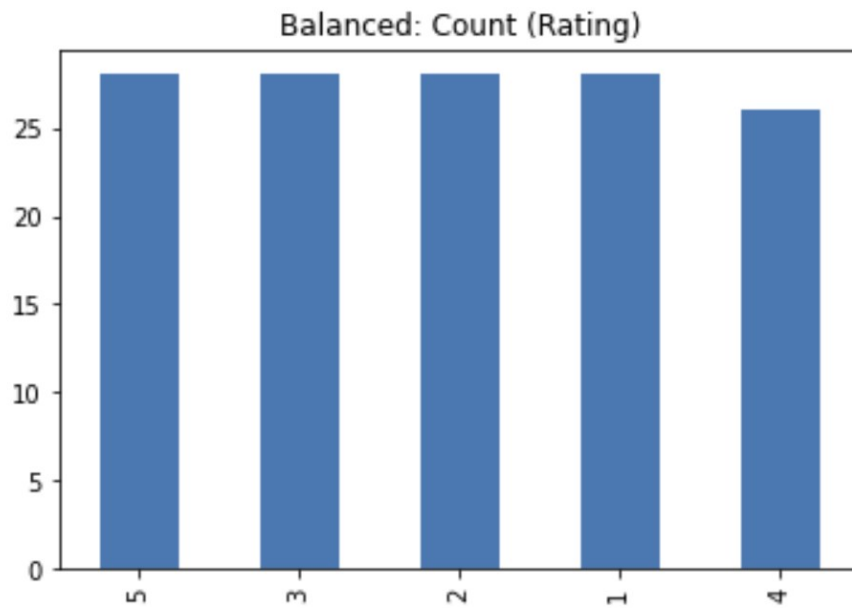
Different features and approaches were experimented with when training a model to predict ratings. A total of 153 samples of facial expressions and ratings were collected. From these 153 samples, 25 samples were found faulty, which left 128 usable samples.

When investigating the data and its distribution, the data was found to have an imbalanced distribution of ratings. The data had more samples where the rating was 4 and 3 than other ratings. This imbalance implied that the model would be highly biased towards ratings that often appeared in the data. Oversampling was applied to adjust for the imbalanced data. Figure 2.7 and figure 2.8 shows the rating distribution before and after applying random oversampling respectively.



(a) Biased towards rating 3 and 4

Figure 2.7: Count distribution of target classes before oversampling



(a) Equally number of targets

Figure 2.8: Count distribution of target classes after oversampling

2.8.3 Features

A number of features and combinations were experimented with as training data. A manual approach and an exhaustive approach was exploited to figure out which features and algorithms to use. The manual approach included visualising the data to make sense of it, while

the exhaustive approach involved iterating through a series of models and feature combinations.

The different feature combinations were extracted from the initial 30 features captured from the web camera. The features are categorized into three groups. Table 2.5 represents these categories. *Expressiveness & Experience* contains emotional degrees such as engagement and valence, which are related to *The Dimensional Model*. *Emotions* represents the individual emotional responses, which are related to *The Universal Emotions Model*. *Facial Expressions* represents each individual nuance of the face. The facial expressions are utilized by Affectiva to derive both emotions and dimensions [1].

Expressiveness & Experience	Emotions	Facial Expressions
Engagement, Valence	Joy, Sadness, Disgust, Contempt, Fear, Surprise, Anger	smile, noseWrinkle, lipPucker, smirk, dimpler, innerBrowRaise, upperLipRaise, lipPress, eyeClosure, eyeWiden, browRaise, lipSuck, lipCornerDepressor, lidTighten, cheekRaise, browFurrow, chinRaise, mouthRaise, mouthOpen, jawDrop, lipStretch

Table 2.5: All features captured with Affectiva API.

Each feature represents a vector of values, where each value is between -100 to 100 for the *Expressiveness & Experience* category, and from 0 to 100 for the *Emotions & Facial Expressions* categories. Each value is the captured facial expression at a specific frame from the web camera. Several single value measurements were extracted from the feature vectors to find the best combination that would explain the data for a prediction algorithm. The calculated values were *Median*, *Mean*, *Standard Deviation*, *Min*, and *Max*.

2.8.4 Models

Several models were trained with a number of feature combinations in search of the best prediction model. For each iteration, different scalers were applied, such as *Standard-scaler*, *MinMax-scaler*, *Robust-scaler* and *MaxAbs-scaler*. The best model from the exhaustive iteration was serialized and used in the *main-study*.

The models selected for this exhaustive search was *Logistic Regression*, *Random Forest Classifier*, *Linear Support Vector Machine*, *Gradient Boosting Classifier* and *K-Nearest Neighbors Classifier*. Figure 2.9 presents an image fragment of trained models, the scaler used, and feature combinations printed in the iterations, while table 2.6 present the best result obtained for each of the models.

```

linear-support-vector-machines
linear-support-vector-machines: 0.372 with standard-scaler and ['_median']
linear-support-vector-machines: 0.349 with minmax-scaler and ['_median']
linear-support-vector-machines: 0.302 with robust-scaler and ['_median']
linear-support-vector-machines: 0.349 with maxabs-scaler and ['_median']
gradient-boosting
gradient-boosting: 0.279 with standard-scaler and ['_median']
gradient-boosting: 0.302 with minmax-scaler and ['_median']
gradient-boosting: 0.302 with robust-scaler and ['_median']
gradient-boosting: 0.302 with maxabs-scaler and ['_median']
logistic-regression
logistic-regression: 0.349 with standard-scaler and ['_median', '_mean']
logistic-regression: 0.349 with minmax-scaler and ['_median', '_mean']
logistic-regression: 0.349 with robust-scaler and ['_median', '_mean']
logistic-regression: 0.349 with maxabs-scaler and ['_median', '_mean']
random-forest

```

(a) Excerpt from the exhaustive model search

Figure 2.9: Best results for each model

Model	Features	Scaler	CV Accuracy Score (Avg)	CV Accuracy Score (Std)	Precision (Macro)	Recall (Macro)	F1 (Macro)
Logistic Regression	Mean	Standard Scaler	0.468	0.063	0.405	0.433	0.368
Random Forest Classifier	Min, Max, Mean, Median	Standard Scaler	0.532	0.060	0.330	0.383	0.330
Linear Support Vector Machine	Mean	Standard Scaler	0.495	0.054	0.468	0.427	0.366
Gradient Boosting Classifier	Min, Max, Std, Median	MinMax Scaler	0.543	0.098	0.418	0.456	0.421
K-Nearest Neighbors	Min, Max, Mean	Robust Scaler	0.439	0.082	0.320	0.454	0.334

Table 2.6: Best prediction model: Gradient Boosting Classifier. CV Accuracy Score: 0.543.

Gradient Boosting Classifier took the longest to train but also yielded overall better results. Each iteration of this model had more stable results and higher accuracy on all the combinations. On each iteration, the models were trained using cross-validation with five folds. The column *CV Accuracy Score* in table 2.6 presents the average and the standard deviation from these folds. The *K-Nearest Neighbors* model did the worst, while the *Gradient Boosting Classifier* did best and was ultimately serialized and used in the main study to predict new ratings.

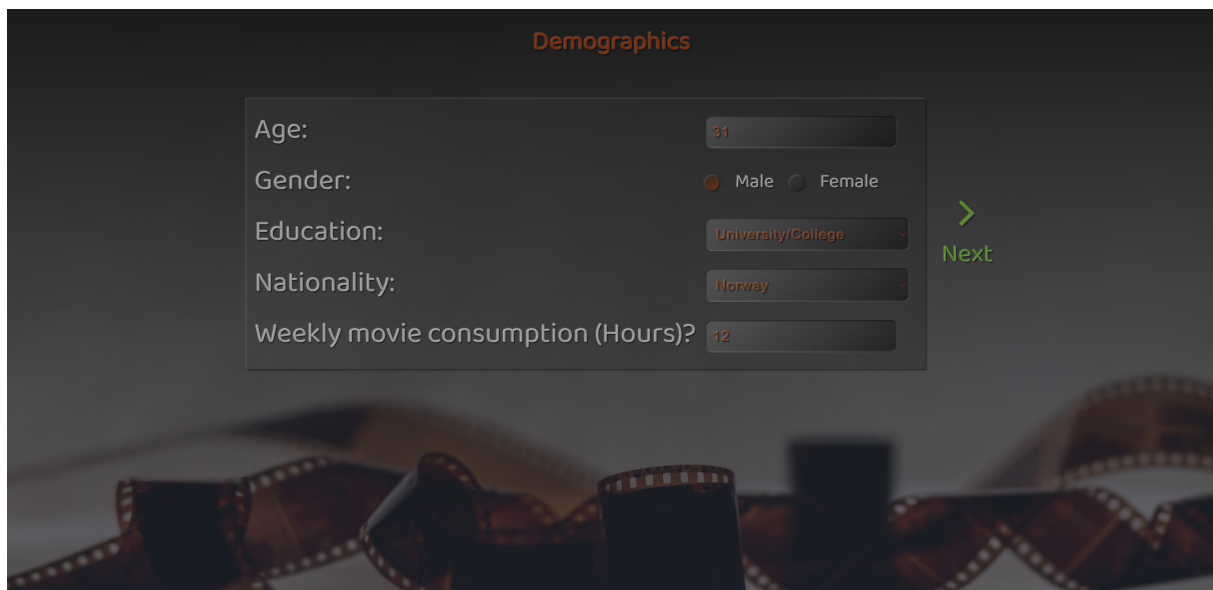
2.9 Main study

This section presents the main study, which includes the user flow of the system and explanations for each step in the experiment. The system design presented here is the complete system, including the prediction model presented in the previous section. Participants in the experiment had to complete each step to continue to the next step, and each step had the purpose of collecting data to build a user profile.

2.9.1 Design

Step 1: Demographics

In image 2.10, we see the first step after creating an account and verifying the web camera. This step asked the user to provide demographic information such as age, gender, education, nationality, and an estimate of how many hours they spend watching movies each week.



The image shows a dark-themed user interface for a 'Demographics' step. The title 'Demographics' is at the top in orange. Below it, there are five input fields: 'Age:' with a text box containing '31'; 'Gender:' with radio buttons for 'Male' (selected) and 'Female'; 'Education:' with a dropdown menu showing 'University/College'; 'Nationality:' with a dropdown menu showing 'Norway'; and 'Weekly movie consumption (Hours)?' with a text box containing '12'. To the right of these fields is a green arrow pointing right, labeled 'Next' in green text. The background of the form is a blurred image of film strips.

Figure 2.10: Demographics

Step 2: Personality Questionnaire

Step 2, depicted in 2.11, presents a personality questionnaire to the user. The questionnaire consists of 10 questions to assess users personalities based on the TIPI: Ten Item Personality Inventory [21].

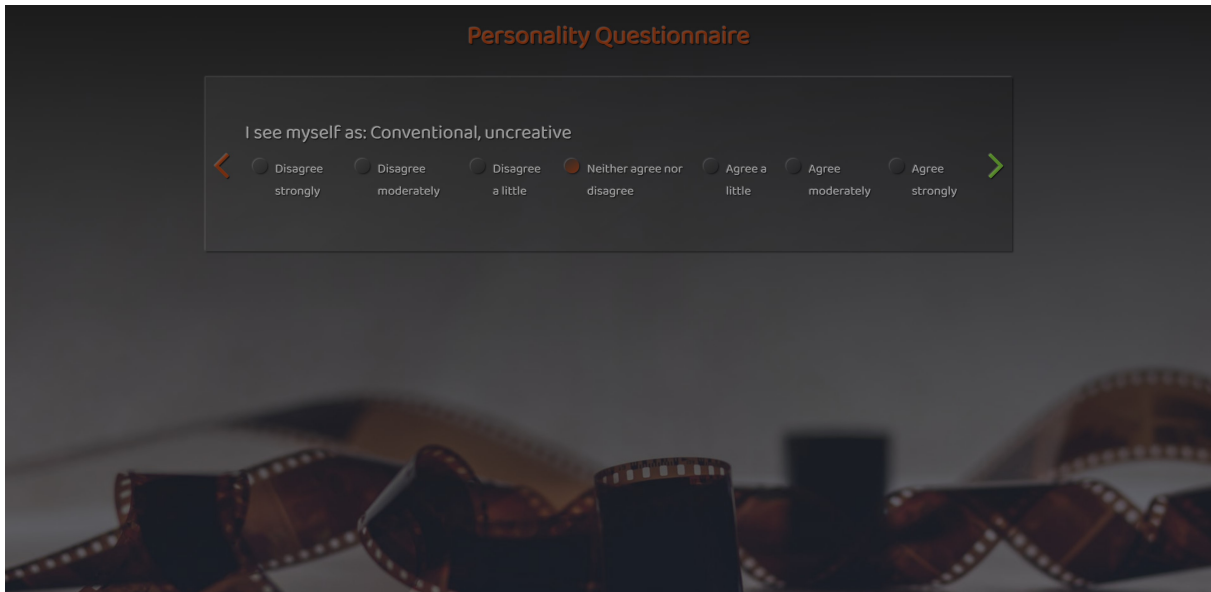


Figure 2.11: Personality Questionnaire

Step 3: Favourite Genre

After collecting demographic data and personality data, the system transitions over to collect high-level movie preferences. In this step, the user select a favorite genre, which the system uses to filter and present a set of movies to watch. Image 2.12 depicts the design and possible genres for the user to select.

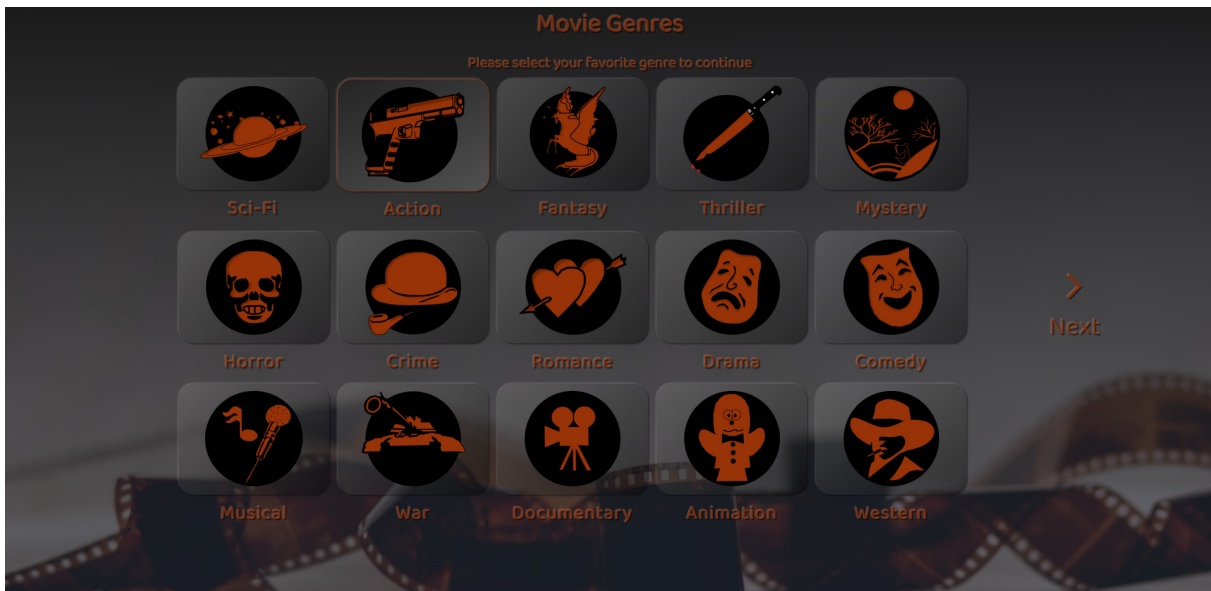


Figure 2.12: Favourite Genre

Step 4: Selecting Movies

After the user selects a genre, the system presents a set of movies. The presented movies are filtered on the previously selected genre. In addition, the movies are sorted by their popularity to get more overlap when collecting facial expressions. We counted the number of ratings for each movie while calculating the average rating. By defining a threshold of 30 counts, movies below this threshold will not be considered popular. This is because a movie can have a high average rating but with a small number of total ratings. Table 2.7 presents a list of the ten most popular movies.

In addition to presenting the user with an initial selection of movies, the user can also shuffle the list to get new candidates, filter by different decades, and create a playlist of a minimum of four movies or more. The design of this window is depicted in image 2.13.

MovieId	Average Rating	Rating Count	MovieTitle
318	4.584615	325	Shawshank Redemption, The (1994)
858	4.527869	305	Godfather, The (1972)
922	4.507692	65	Sunset Blvd. (a.k.a. Sunset Boulevard) (1950)
1198	4.501340	373	Raiders of the Lost Ark (Indiana Jones and the Raiders of the Lost Ark) (1981)
1223	4.491803	61	Grand Day Out with Wallace and Gromit, A (1989)
1193	4.489451	237	One Flew Over the Cuckoo's Nest (1975)
260	4.464126	446	Star Wars: Episode IV - A New Hope (1977)
2357	4.459459	37	Central Station (Central do Brasil) (1998)
904	4.455172	145	Rear Window (1954)
2324	4.454545	165	Life Is Beautiful (La Vita è bella) (1997)

Table 2.7: Top 10 popular movies from all genres and years

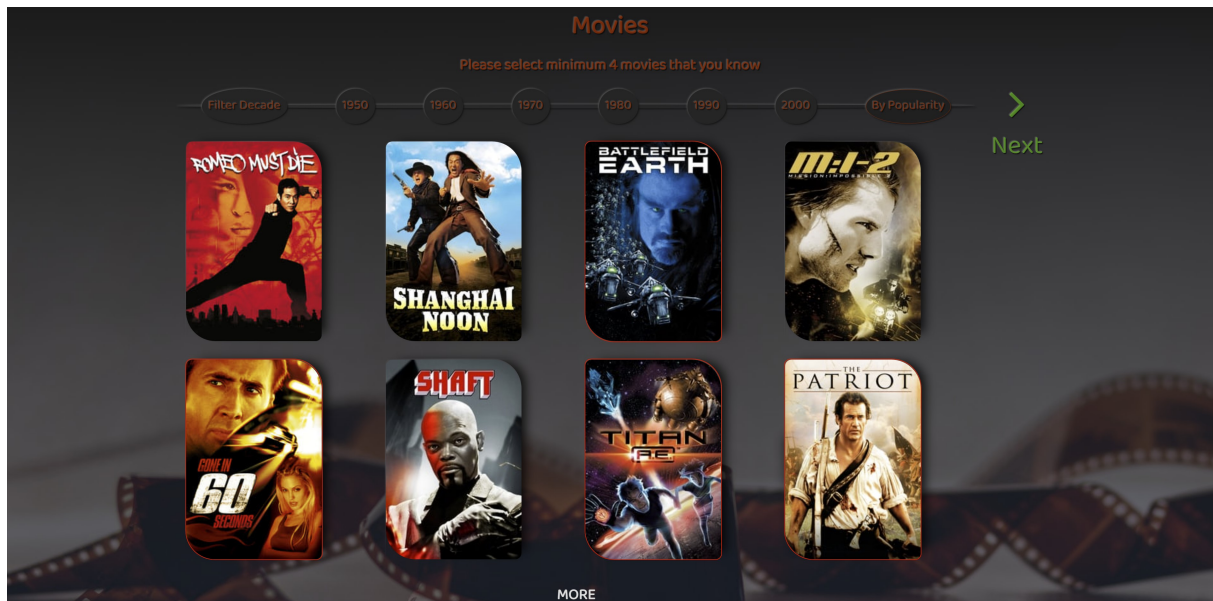


Figure 2.13: Selecting Movies

Step 5: Watching and Rating

In step 5, the user enters the consumption stage explained by Tkalčič et al. [42]. The user watches all of the selected movie trailers in sequence. For each trailer, the system collects facial expressions and asks the user to provide an explicit rating. The user continues to the next movie trailer in the playlist after completing the task. In this step, the system builds the affective user profile, which is used to predict ratings. The design of this step is depicted in image 2.14.

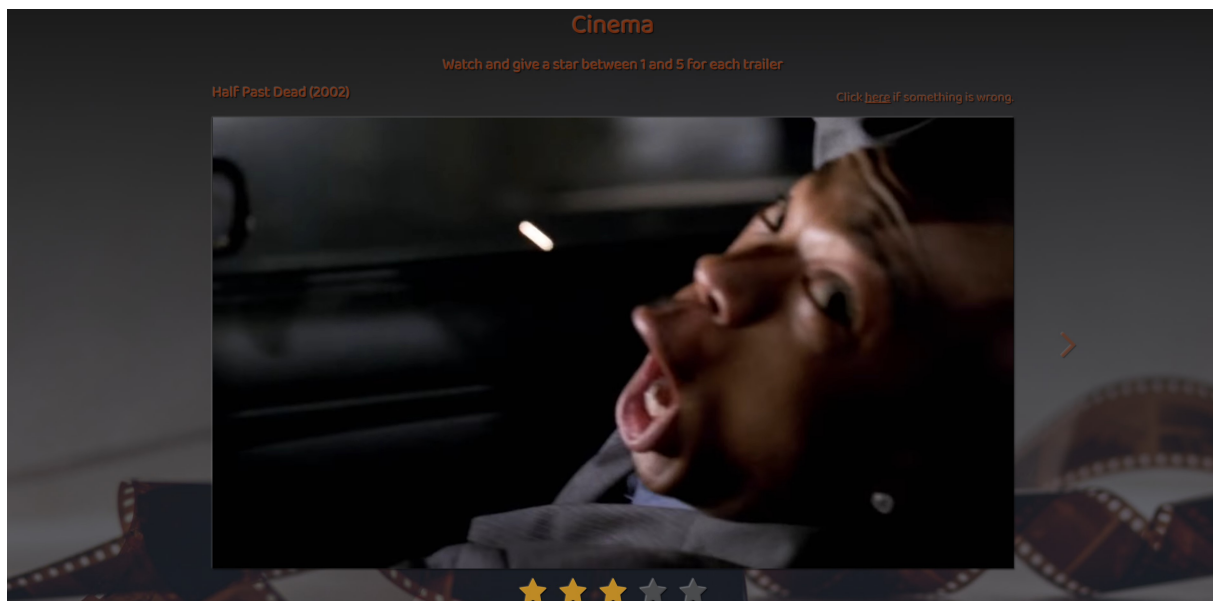


Figure 2.14: Watching and Rating

Step 6: Recommendations

The next step after watching movie trailers, the user navigates to the recommendation window. This window is depicted in image 2.15, and present the user with three lists of recommendations. One list is generated using the Emotion-based Filtering Technique. Another is generated using Collaborative Filtering (CF). And a third list is generated using Visual-based Filtering Technique.

In this step, the user evaluate three lists based on four questions. The first two questions aim to understand the *Accuracy*, while the last two questions aims at understanding *Diversity*. The user answers each question by selecting one of the lists after comparing the recommendation lists. When all the questions are answered, the user continues to the last step.

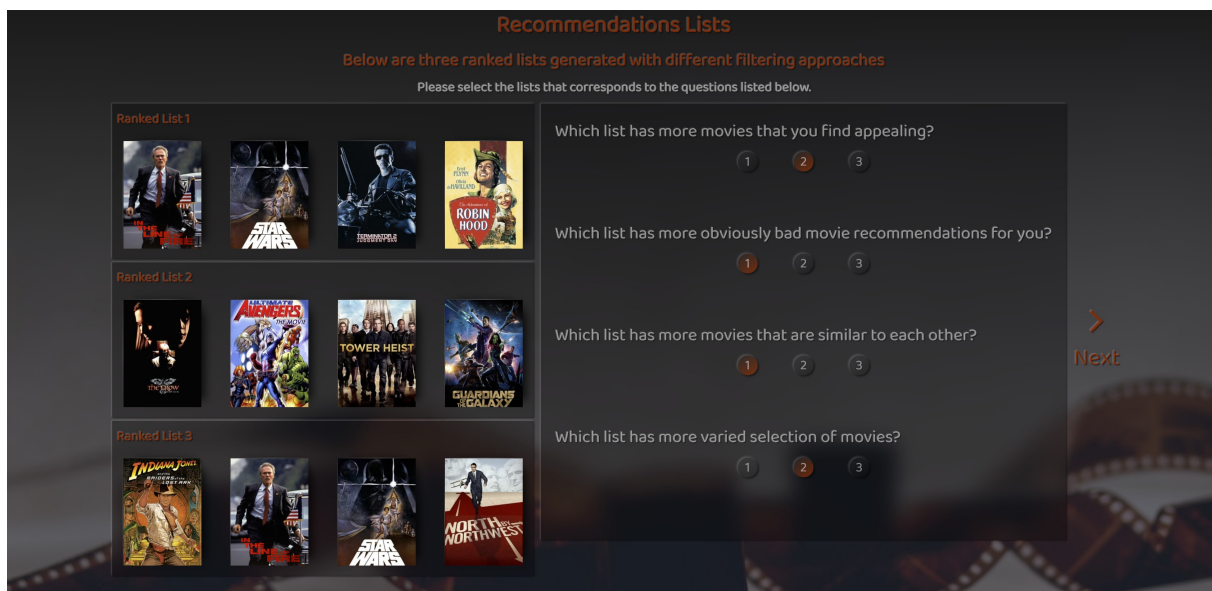
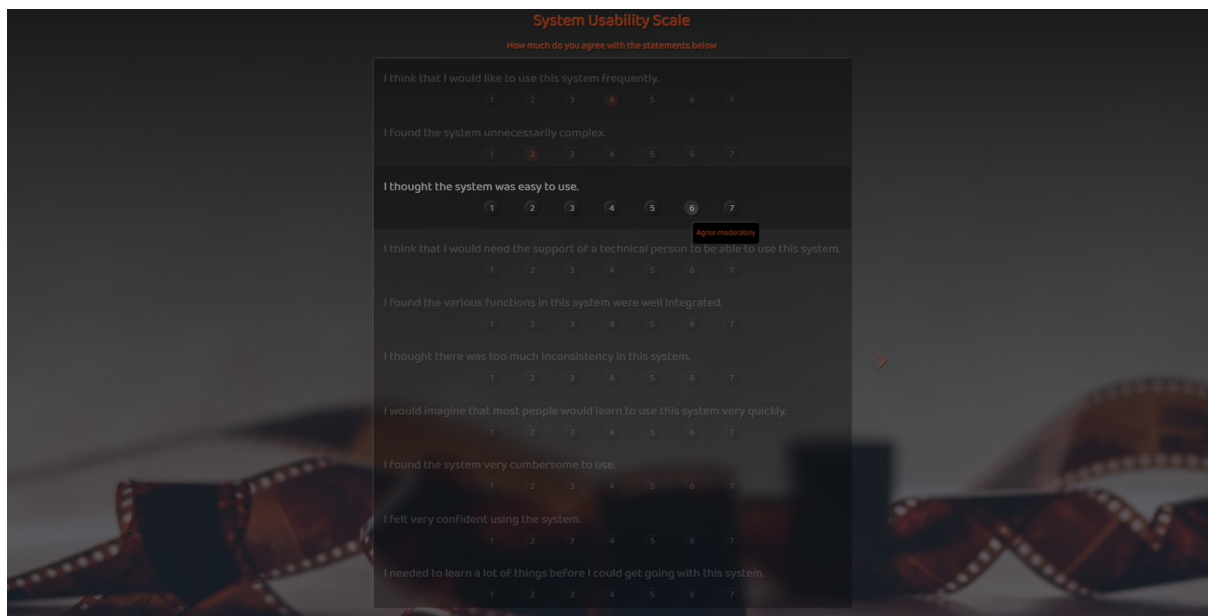


Figure 2.15: Recommendations

Step 7: System Usability Questionnaire

The last step in the experiment presents the user with a System Usability Questionnaire (SUS) depicted in 2.16. This step aims to understand the system's usability by agreeing or disagreeing with a set of ten statements. When completing this step, the user is presented with a message which thanks for the participation to make sure the user knows the experiment is completed.



The image shows a screenshot of a System Usability Scale (SUS) questionnaire. The title is "System Usability Scale" in orange. Below it, the instruction reads "How much do you agree with the statements below". The questionnaire consists of ten statements, each followed by a horizontal scale from 1 to 7. The scales are as follows:

- Statement 1: "I think that I would like to use this system frequently." Scale: 1, 2, 3, 4, 5, 6, 7. The number 4 is highlighted in red.
- Statement 2: "I found the system unnecessarily complex." Scale: 1, 2, 3, 4, 5, 6, 7. The number 2 is highlighted in red.
- Statement 3: "I thought the system was easy to use." Scale: 1, 2, 3, 4, 5, 6, 7. The number 6 is highlighted in red.
- Statement 4: "I think that I would need the support of a technical person to be able to use this system." Scale: 1, 2, 3, 4, 5, 6, 7. The number 6 is highlighted in red, and a tooltip "Agree moderately" is visible above it.
- Statement 5: "I found the various functions in this system were well integrated." Scale: 1, 2, 3, 4, 5, 6, 7.
- Statement 6: "I thought there was too much inconsistency in this system." Scale: 1, 2, 3, 4, 5, 6, 7.
- Statement 7: "I would imagine that most people would learn to use this system very quickly." Scale: 1, 2, 3, 4, 5, 6, 7.
- Statement 8: "I found the system very cumbersome to use." Scale: 1, 2, 3, 4, 5, 6, 7.
- Statement 9: "I felt very confident using the system." Scale: 1, 2, 3, 4, 5, 6, 7.
- Statement 10: "I needed to learn a lot of things before I could get going with this system." Scale: 1, 2, 3, 4, 5, 6, 7.

Figure 2.16: System Usability Questionnaire

2.10 Shortcomings

In the process of developing the artifact, a few shortcomings were noted. One shortcoming was the lack of support for participating in the experiment through tablets and phones. The reason for this constraint was due to time scope and web camera technicalities. Because the interface did not scale to small screens, the system was unusable on small devices. These shortcomings were often the reason that a percentage of users only created an account without proceeding with the experiment.

In addition to not scaling to small devices, the system experienced challenges when scaling for massive usage. This was due to the YouTube API quota limit. If too many participants watched trailers in one day, the quota limit was quickly reached. To handle this, we had to gradually ask people to participate.

The last shortcoming revolved around missing knowledge about the recommended movies. When evaluating the three lists of recommendations, the participants could get more information about a movie recommendation by clicking on the movie. When clicking on a movie, the genre and plot of the movie was presented. Even though this provides additional knowledge to help assessment, we can not exclude the assumption that users can find it difficult to evaluate the recommendation lists with little knowledge about the movies.

2.11 Advantages

While there was noted some shortcomings, a few advantages were also noted. The system was deployed online which made it possible for participants to do the experiment whenever they had time. The online evaluation also made it easier to reach more participants. In addition, an online evaluation makes participants feel freer to navigate the system. This can contribute to a less biased end-result compared to offline evaluation, where participants might tend to evaluate the system on behalf of expectations.

The system design contained all the steps needed to execute and assess the entire experiment. The participants could complete the experiment in about 20 minutes without needing further instructions or supplementary surveys. All of the data was collected and processed within the system, making it easier for people to participate and finish. In addition, the obtained data becomes easier to aggregate, process, and study, as it is stored and structured from a single source.

Chapter 3

Results

This chapter focuses on the online experiment with real users and presents the obtained results. This experiment have been designed to address the research questions:

- **RQ1:** Are there any difference among movie genres in terms of the emotional responses obtained from facial expressions of the users?
- **RQ2:** Is there a correlation between visual features encapsulated within movies and the emotional responses the users express?
- **RQ3:** In terms of Accuracy and Diversity, what is the quality of recommendation based on emotional responses, using facial expressions, in comparison to the other approaches?
- **RQ4:** In terms of Accuracy and Diversity, do users with similar personality traits prefer similar recommendation approaches?
- **RQ5:** Can the preferences of users be elicited from their emotional responses extracted from the facial expressions in order to generate movie recommendations?

The first part presents the results of a preliminary analysis of the obtained user data. The second part presents the results obtained from analysing users emotional responses in movie genres. Then, the results from analysing the correlation between emotional responses and visual features are presented. Further on, results are presented from the online experiment and how users evaluated the recommendation lists in terms of *Accuracy* and *Diversity*. The last section presents the results from the usability study based on the *System Usability Survey* (SUS).

3.1 Users

The experiment had in total 77 people who created an account, 43 of those who completed the whole experiment, and 34 who created an account but did not complete the experiment or only did some of the steps. The distribution of participants in terms of demographics background is presented below. Figure 3.1 represents the distribution of males and females and figure 3.2 shows the distribution of age. The ages are grouped into ranges. Figure 3.3 represents the distribution of education, while figure 3.4 show the distribution of nationality. The hyphen in the nationality figure 3.4 represents a non-disclosure choice.

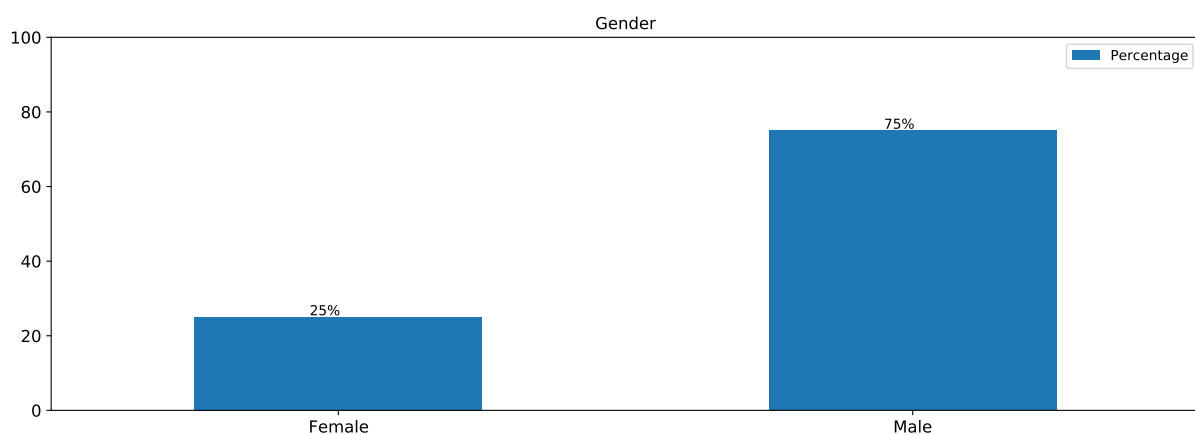


Figure 3.1: Distribution by Gender

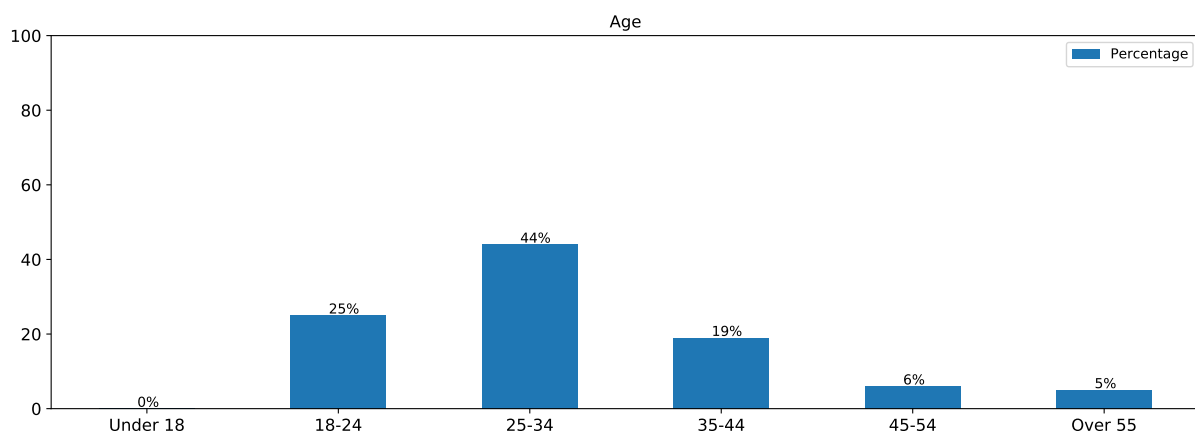


Figure 3.2: Distribution by Age

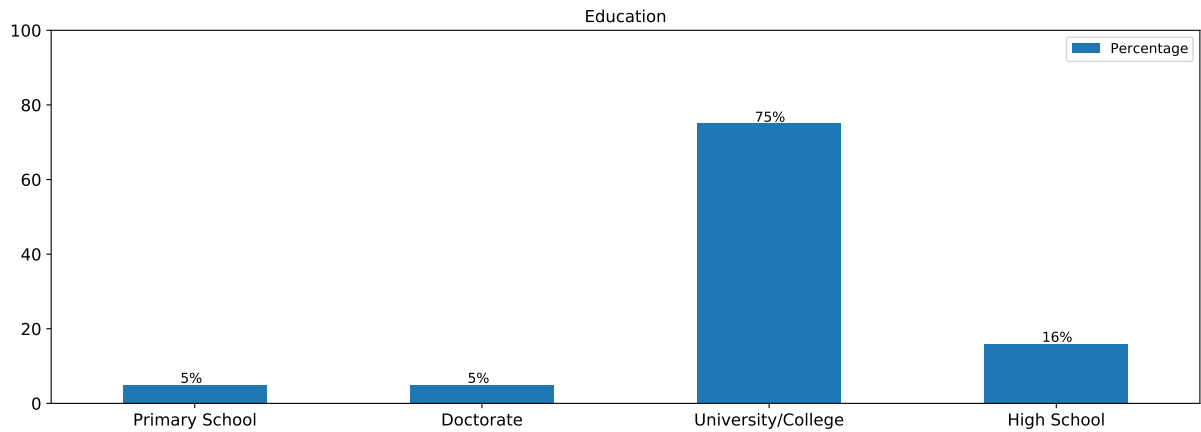


Figure 3.3: Distribution by Education

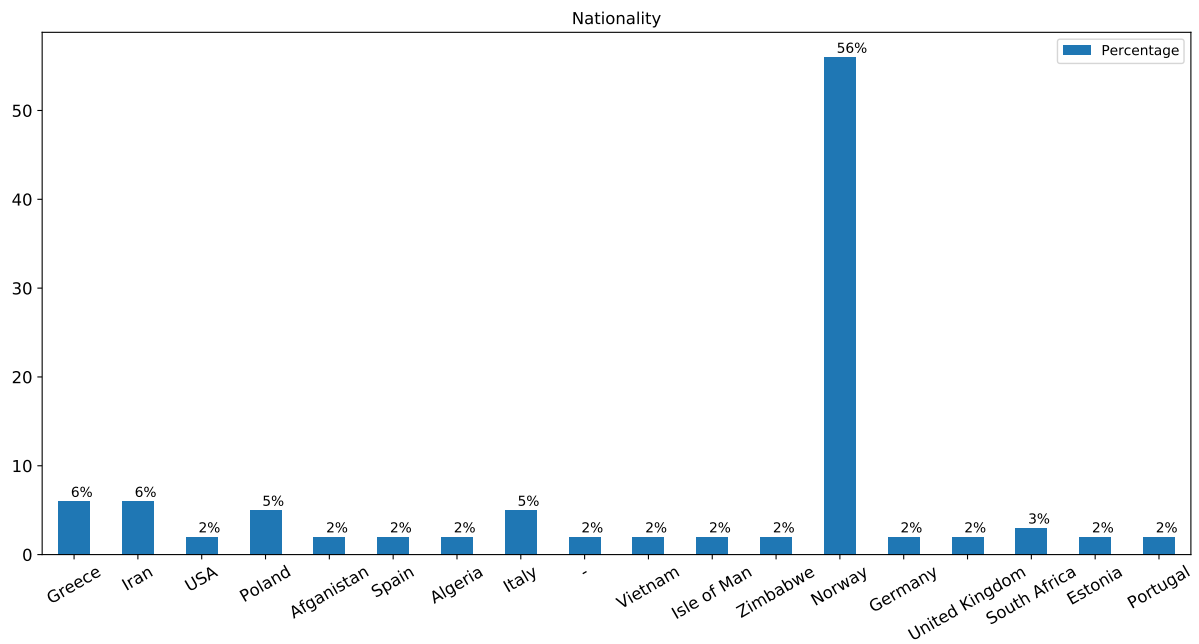


Figure 3.4: Distribution by Nationality

3.2 Emotion & Visuals

3.2.1 Procedure

In order to investigate the potential correlation between emotional responses of users and visual features of the movies, two datasets were been used, i.e., a primary dataset that contains the facial expressions, collected through the experiment and a secondary dataset that

contains the visual features, which had been collected and generated previously by Elahi et al. [14].

Each of the datasets can be represented as a matrix where the columns show the data points for each frame captured. In the dataset of visual features, these frames are key-frames throughout a trailer and a corresponding feature value. The same structure is used in the dataset of facial expressions and emotions. The emotional responses *anger*, *contempt*, *disgust*, *fear*, *joy*, *sadness* and *surprise* has been selected to be analyzed in this study. In order to compare and analyse the movies in the two datasets, the two datasets were merged.

The frames of each movie were encapsulated into chunks of 100 units so that each unit was the average of n frames. The average emotional response and the average visual feature throughout a movie trailers timeline could then be explored and analyzed. When studying the data, features were found to fluctuate across the timeline. This fluctuation was reduced by moving the average using Exponential Moving Average (EWM¹). Figure 3.5 demonstrate the procedures taken to prepare the data into units, while figure 3.6 demonstrates the fluctuation on the features *contrast* and *joy* on a random movie before and after moving the average.

¹<https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.ewm.html>

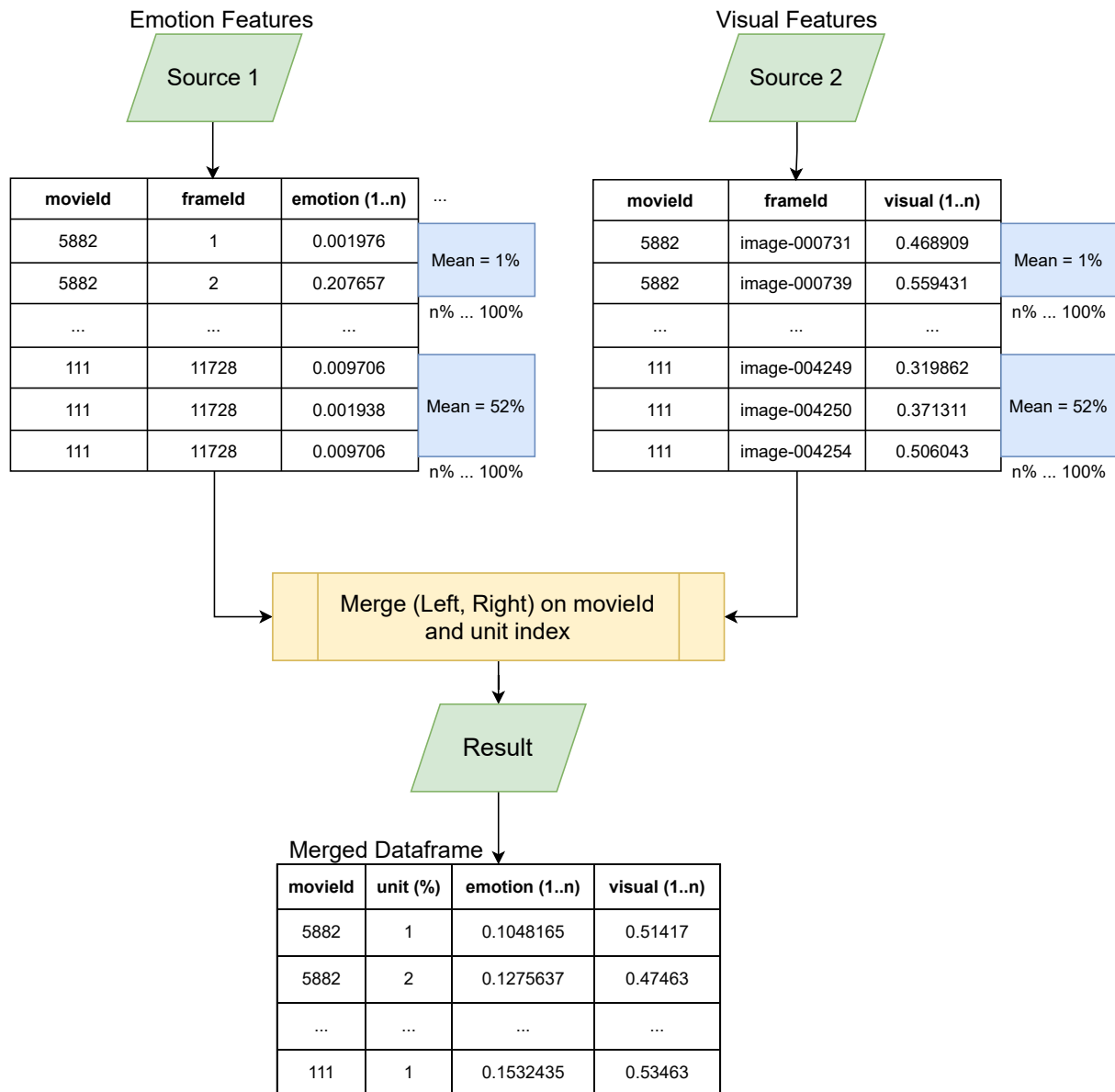


Figure 3.5: Datasets merged into units: Each movie in the new dataframe has units going from 1 to 100

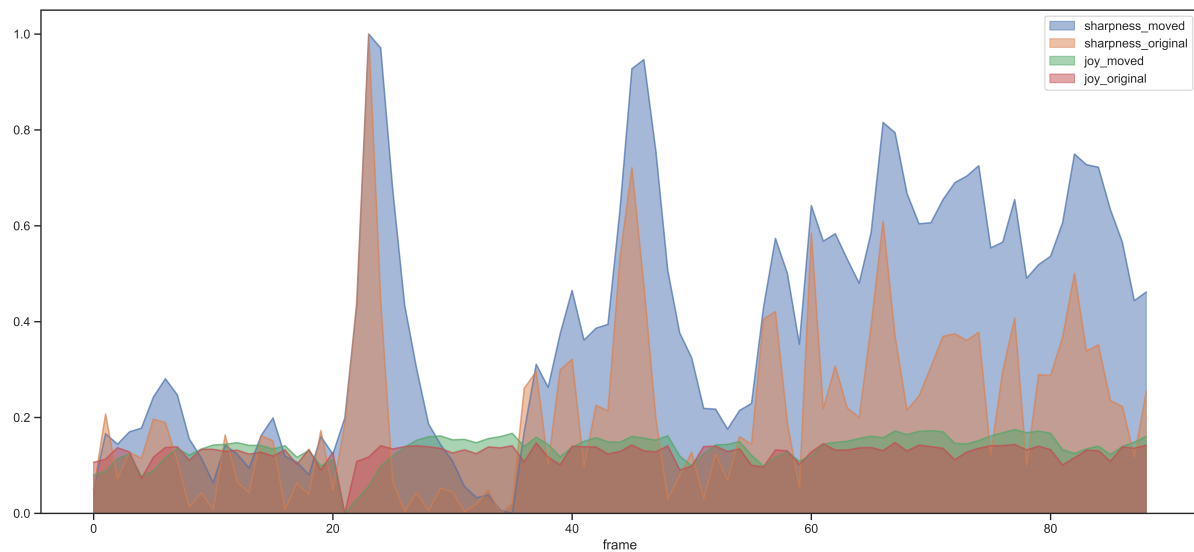


Figure 3.6: Fluctuation of the feature *contrast* and *joy* on a random movie before and after moving the average.

3.2.2 Average Emotion

This section report the results of the analysis aimed at finding the average emotion captured from facial expressions in each genre. This has been performed in addressing the research question:

RQ1: *Are there any difference among movie genres in terms of the emotional responses obtained from facial expressions of the users?*

Figure 3.7 presents the number of movie trailers within each genre. The average emotion was calculated across all movies within each genre.

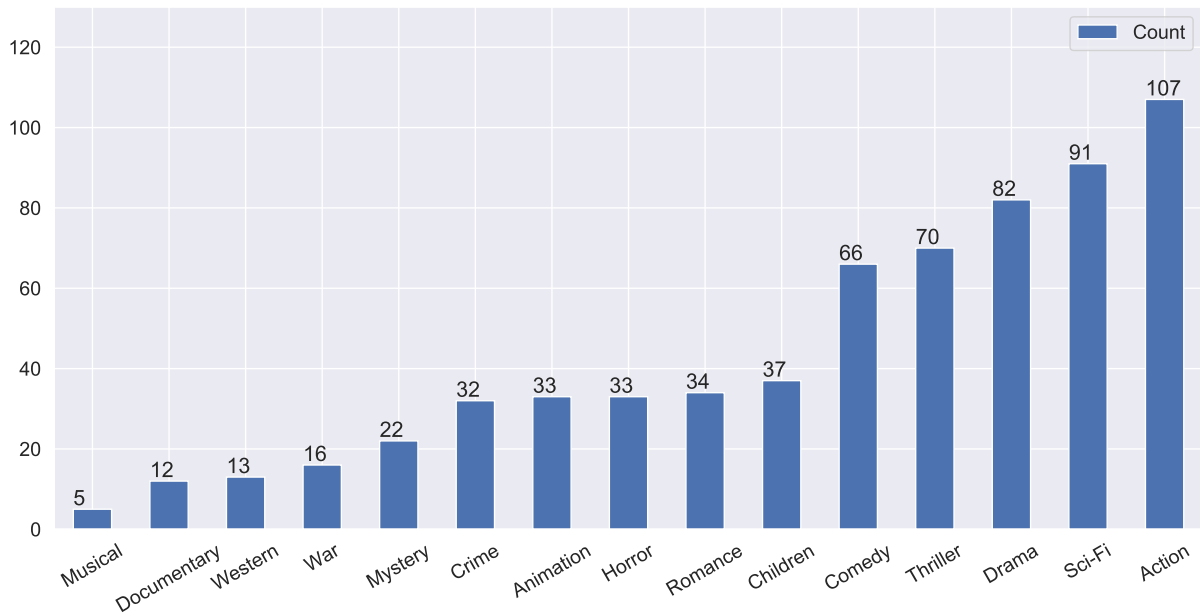


Figure 3.7: Count of movies within each genre

Figure 3.8 is a matrix which presents the average emotion for each movie genre. The emotions follow the x-axis, while the movie genres follow the y-axis. In addition to showing the average emotion, the matrix is decorated with two dendrograms. The top dendrogram represents the hierarchical relationship between emotions, while the left dendrogram represents the hierarchical relationship between movie genres calculated by the average correlation.

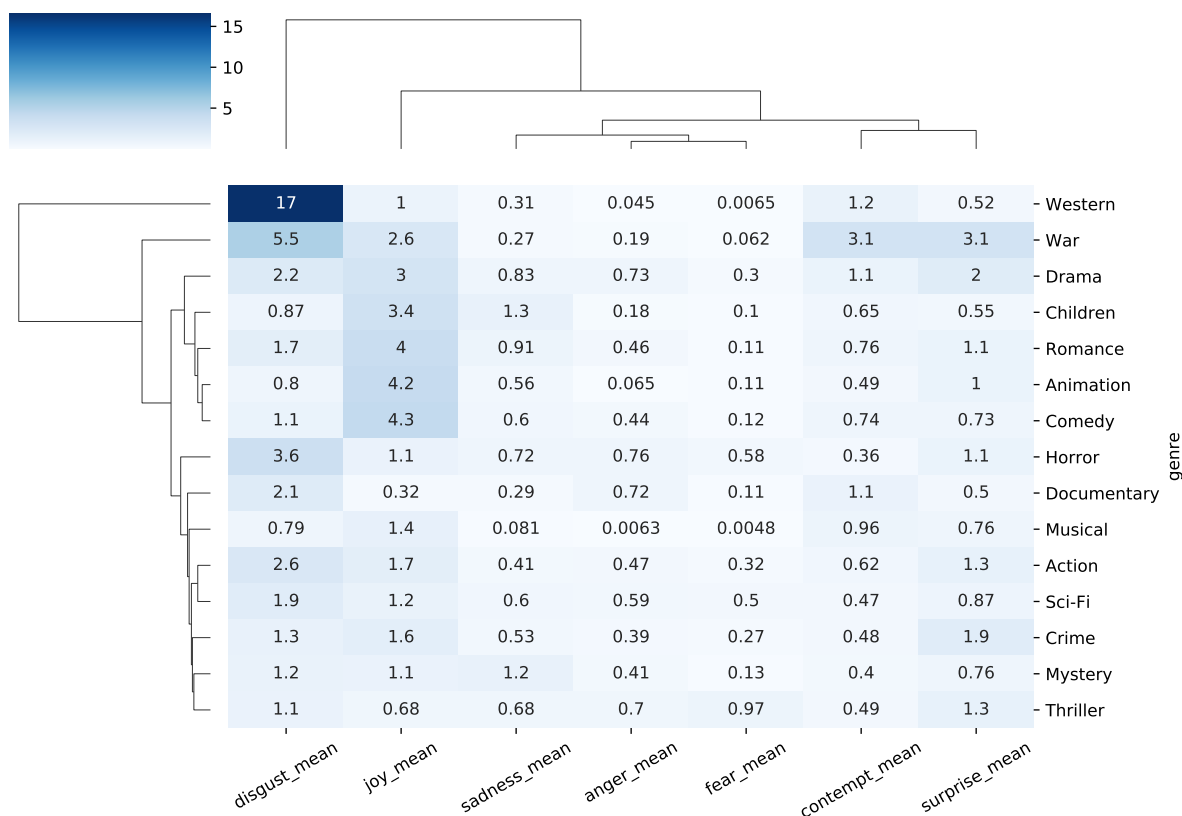


Figure 3.8: Genres average emotion

From the matrix, we can observe that the overall average emotion measured in the genres is low. The highest emotional responses was found in the genres *Western* and *War*. Both of the genres were among the least watched genres and induced the emotional response *disgust*. While there is no clear indication that each genre induce a specific emotional response, we found *disgust* and *joy* to be the most prominent across the genres. Another notable observation is that genres such as *Comedy*, *Romance*, *Animation* and *Children* have a higher average emotion of *joy* than genres such as *Horror*, *Western*, *War*, *Action* and *Sci-Fi*, which tend to induce more of *disgust*.

In contrast to analysing distinct differences in average emotion between genres, the dendrograms display the similarities. From the top dendrogram, we can see that *fear* and *anger* are the most similar across genres. Close to *fear* and *anger*, we see that *contempt* and *surprise* are similar across the genres. In addition to the emotional similarities across genres, the left dendrogram shows clusters of similar genres. The most similar genres are *Animation* and *Comedy*, *Action* and *Sci-Fi*, and *Mystery* and *Thriller*.

3.2.3 Correlation

This section reports the results of the analysis aimed at finding potential correlations between emotional responses and visual features. This has been performed in addressing the research question:

RQ2: *Is there a correlation between visual features encapsulated within movies and the emotional responses the users express?*

To compute the correlation between visual features and the emotional responses from users, each movie was transformed into a correlation matrix. The resulting matrices contained the correlations between visual features and emotional responses based on the average variation captured from units of key-frames. The matrices were further grouped by genre to average the correlations within the same population. Each correlation value was first transformed to Fisher's Z-Scores before averaging the values. By normalizing the sampling distribution of correlations using Fisher's Z transformation, we can average the Z-scores, and then back-transform the averaged values to correlation coefficients. As reported by Dunlap et al. [10] and Silver and Dunlap [39], this process reduces the affect of sampling distribution skew, and ultimately results in less biased statistic. In figure 3.9 we present the process of averaging correlations.

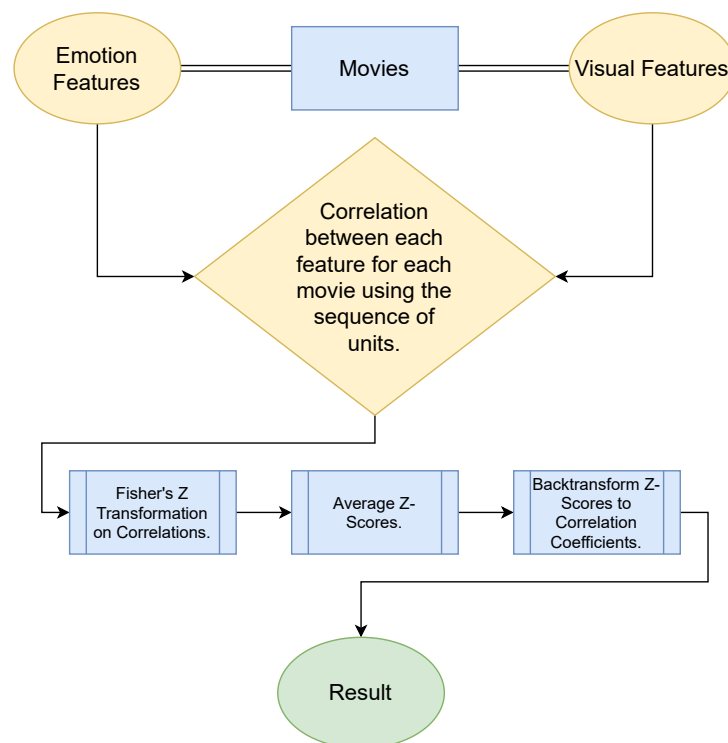


Figure 3.9: Steps taken to find relationship between emotional responses and visual features

Figure 3.10 show the correlation distribution for each visual feature and emotional response.

The y-axis represents the density, while the x-axis represents the correlation coefficients. Each line in the plot shows the emotional response with a corresponding color which can be observed in the legend. The thickness of line represents the average correlation found in the distribution. This means that thicker lines have a higher average correlation than thinner lines.

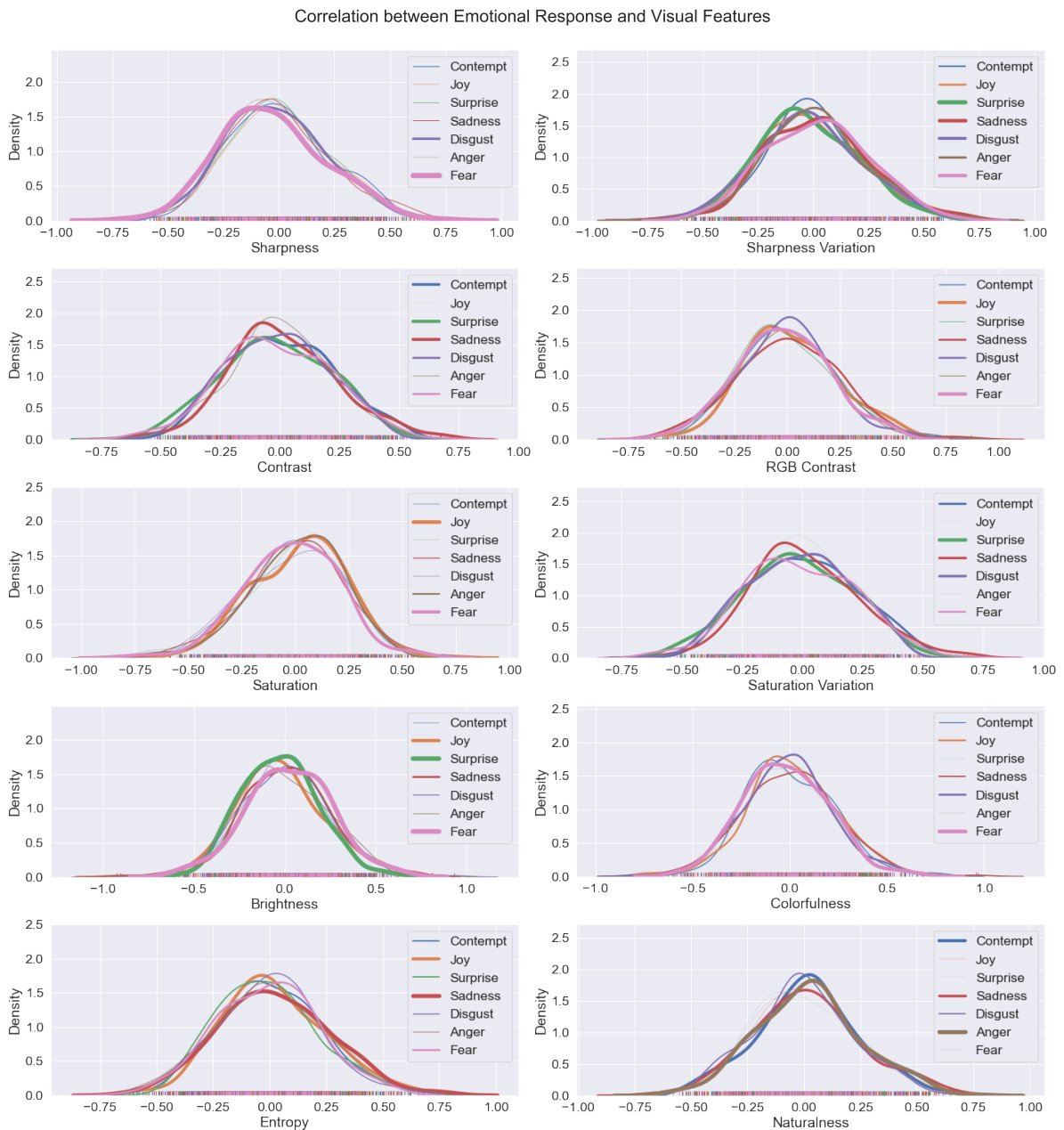


Figure 3.10

For observing the distributions, we see that correlations range from around -0.75 to 0.75, while the density peaks between -0.25 and 0.25. From calculating the average correlation, the top three correlations for each visual feature was selected as potential correlations in table 3.1. The findings are represented with a number to indicate the order.

	Contempt	Joy	Sadness	Disgust	Fear	Anger	Surprise
Sharpness	3			2	1		
Sharpness Variation			2	3			1
Contrast	3		2				1
RGB Contrast		2		3	1		
Saturation		1			2	3	
Saturation Variation			3	2			1
Brightness		3			2		1
Colorfulness		3		2	1		
Entropy		2	1		3		
Naturalness	2		3			1	
Count (1)	0	1	1	0	3	1	4
Count (2)	1	2	2	3	2	0	0
Count (3)	2	2	2	2	1	1	0
SUM	3	5	5	5	6	2	4

Table 3.1: Top three potential correlations (1 = strongest avg. correlation, 2 = second strongest avg. correlation, 3 = third strongest avg. correlation). Count: Sum of each potential correlation.

From observing table 3.1, *Joy* seems to have a potential correlation with *Saturation*. *Sadness* seems to have a potential correlation with *Entropy*. *Anger* seems to have a potential correlation with *Naturalness*. *Fear* and *Surprise* has potential correlation with multiple visual features. *Fear* was found to have potential correlation with *Sharpness*, *RGB Contrast*, and *Colorfulness*. *Surprise* was found to have potential correlation with *Sharpness Variation*, *Contrast*, *Saturation Variation*, and *Brightness*.

In addition to calculating the correlation distributions, the average correlations within each genre were calculated. Figure 3.11 shows a grid of correlation matrices for each genre with the average correlations per feature. Table 3.2 displays the strongest correlation from each genre.

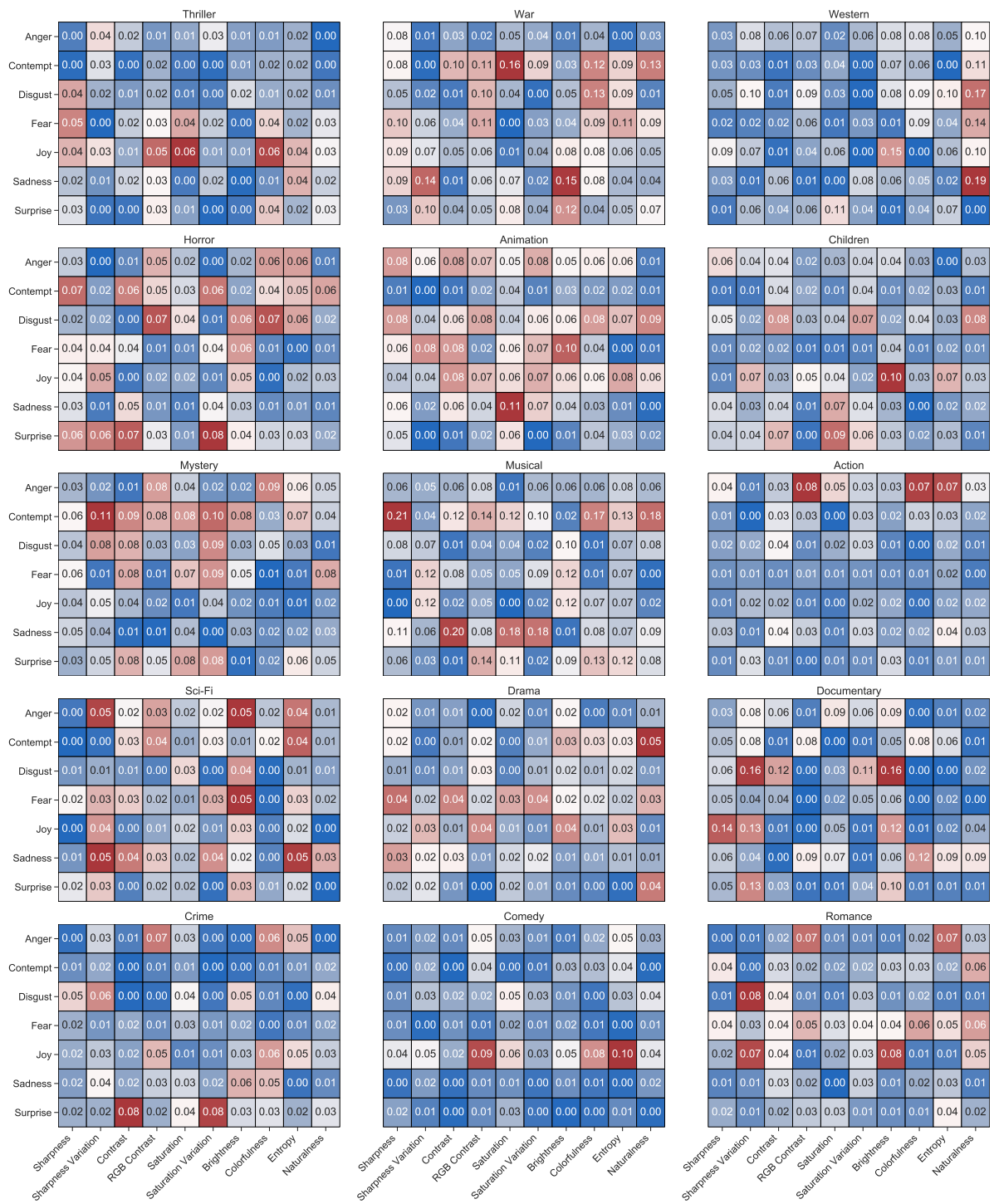


Figure 3.11: Correlation Matrix for Each Genre (Absolute values)

Genre	Emotional Response	Visual Feature	Correlation (r)
Musical	Contempt	Sharpness	0.21
Western	Sadness	Naturalness	0.19
War	Contempt	Saturation	0.16
Documentary	Disgust	Brightness	0.16
Animation	Sadness	Saturation	0.11
Mystery	Contempt	Sharpness Variation	0.11
Comedy	Joy	Entropy	0.10
Children	Joy	Brightness	0.10
Action	Anger	RGB Contrast	0.08
Horror	Surprise	Saturation Variation	0.08
Crime	Surprise	Contrast	0.08
Romance	Disgust	Sharpness Variation	0.08
Thriller	Joy	Saturation	0.06
Sci-Fi	Anger	Brightness	0.05
Drama	Contempt	Naturalness	0.05

Table 3.2: Correlations per genre (Absolute values)

From observing table 3.2, correlations were found to be relatively weak when aggregating by genre. Correlations also tend to get weaker as a genre contains more watched movies. While the correlations are weak, the summary provides an indication of potential correlations for movies in a genre. For example, *Comedy* seems to have potential correlations between *Joy* and *Entropy*, while *Horror* seems to have movies where *Surprise* correlates with *Saturation Variation*. The strongest correlation was found in *Musical*, with a correlation of 0.21 between *Contempt* and *Sharpness Variation*.

3.3 Recommendation Evaluation

This section describe the results of the experiment where we were interested to investigate the performance of the implemented recommendation techniques. Different recommendation approaches was implemented, i.e., one approach was Emotion-based Filtering (EF), the baseline approach was Collaborative Filtering (CF), and the third was Visual-based Filtering (VF).

First, a brief explanation of how users evaluated the recommendation lists, and how similar personalities was clustered. Secondly, the obtained results from all the participants without clustering them by personality traits is presented. Finally, the obtained results from the recommendation evaluation by each group of similar personalities is presented.

3.3.1 Procedure

The steps taken to evaluate the recommendation approaches was conducted through a questionnaire. The questionnaire was given to each user after they had watched minimum four or more movie trailers. Together with the questionnaire, they were given three lists which contained four movie recommendations each. The recommendation approaches generated one list each.

In the questionnaire, the user was asked to compare the lists and select one of the three lists for each question. The questionnaire contained four questions where two of them aimed at evaluating the *Accuracy*, and two of them aimed at evaluating *Diversity*. No information about which list conformed to which recommendation approach were given to the participants. The following presents the questions each participant were asked:

- Accuracy:
 1. Which list has more movies that you find appealing?
 2. Which list has more obviously bad movie recommendations for you?
- Diversity:
 1. Which list has more movies that are similar to each other?
 2. Which list has more varied selection of movies?

In presenting the obtained results from the recommendation evaluation, each evaluated question is presented separately as they are formulated as positive and negative questions. Obtained results which addresses *Accuracy* is presented first, then the results which addresses *Diversity* is presented. Each question show the percentage of how many votes a recommendation technique obtained from all the participants.

After presenting the overall user evaluation, the results obtained from clustering groups with similar personality traits are presented. The big five personality traits of participants were calculated from a questionnaire with 10 statements (called TIPI: Ten Item Personality Inventory) [21]. The participants evaluated each statement by selecting how much they agreed with it. The scale for each statement goes from 1 (disagree strongly) to 7 (agree strongly). Each statement assessed a personality trait of *extraverted*, *critical*, *dependable*, *anxious*, *openness*, *reserved*, *sympathetic*, *disorganized*, *calm* and *conventional*. For the latter, the scores were reversed based on the observed answers of the negative questions and averaged on each corresponding score obtained for the positive questions.

Further on, *Principal Component Analysis* was used to find the two most important component that could be used to cluster participants with similar personalities. Before applying

clustering, *The Elbow Method* was used to find the optimal k clusters in the data. With the Elbow method, four clusters were found to be optimal number of personality clusters.

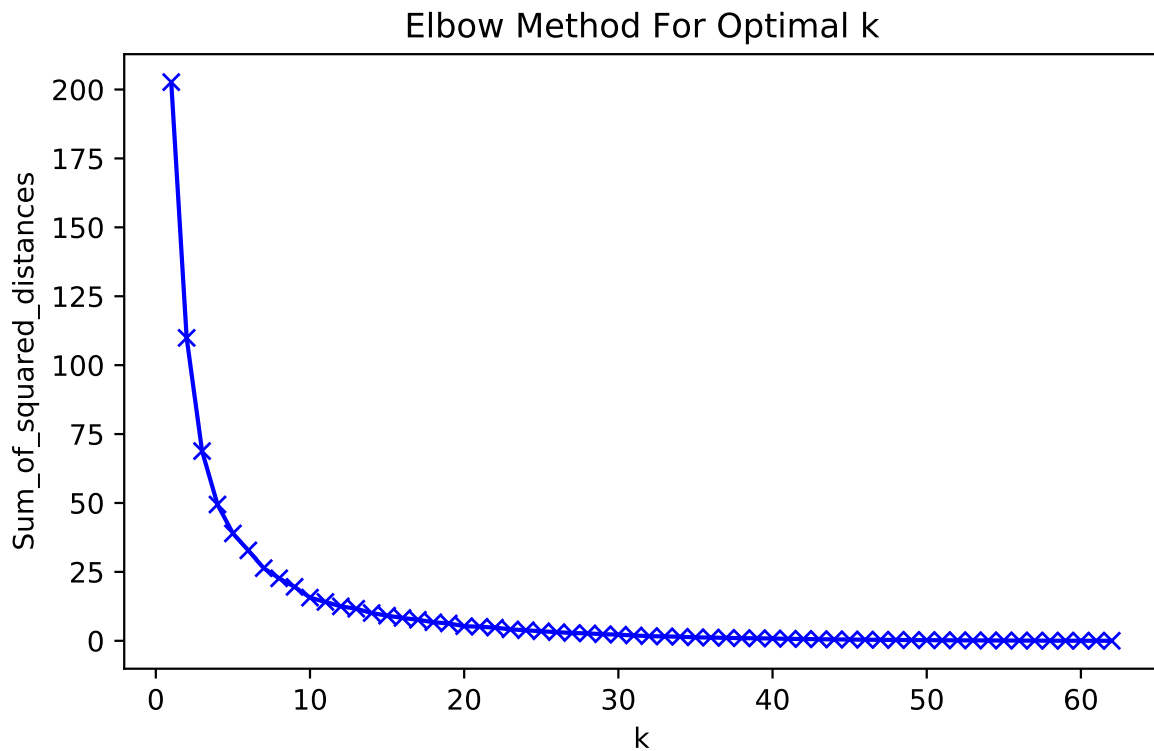


Figure 3.12: Optimal K Clusters.

Figure 3.12 shows the elbow for finding the optimal k clusters. When the optimal k was selected, *KMeans* clustering was used to group the participants with similar personalities. Figure 3.13 presents the distribution of the 4 clusters where each datapoint is based on the first two principal components found using the big five personality traits of each participant.

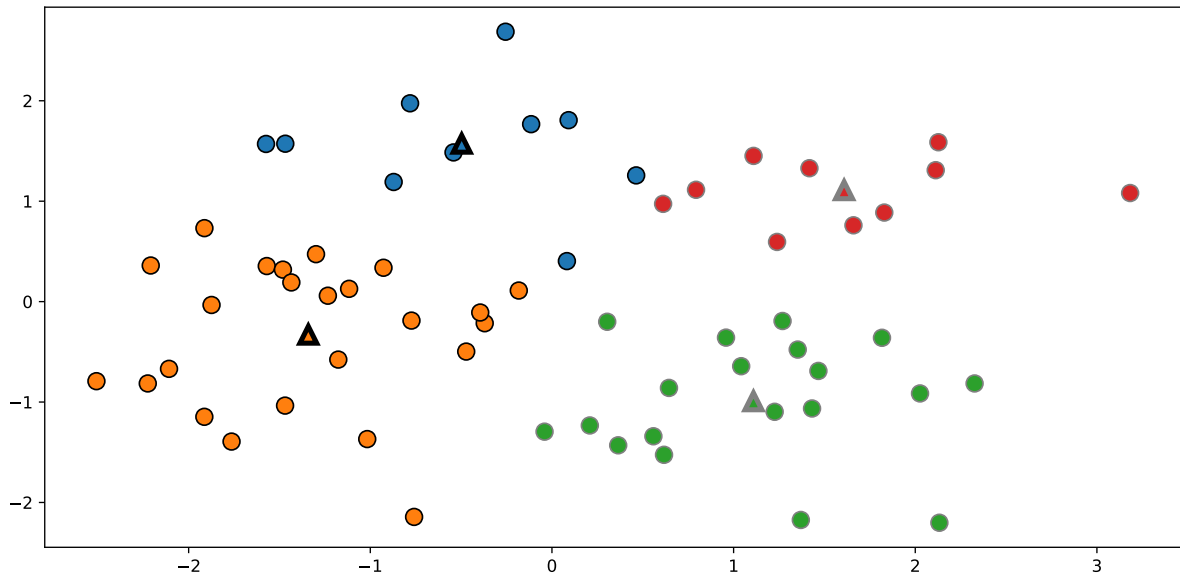


Figure 3.13: Clustered participants with similar personalities

The average personality trait across all participants and the average personality trait within each cluster were calculated to further understand the personalities in each cluster. Figure 3.14 presents the clusters where the bars represent the average personality traits in each cluster, while the lines represent the average personality traits across all participants.

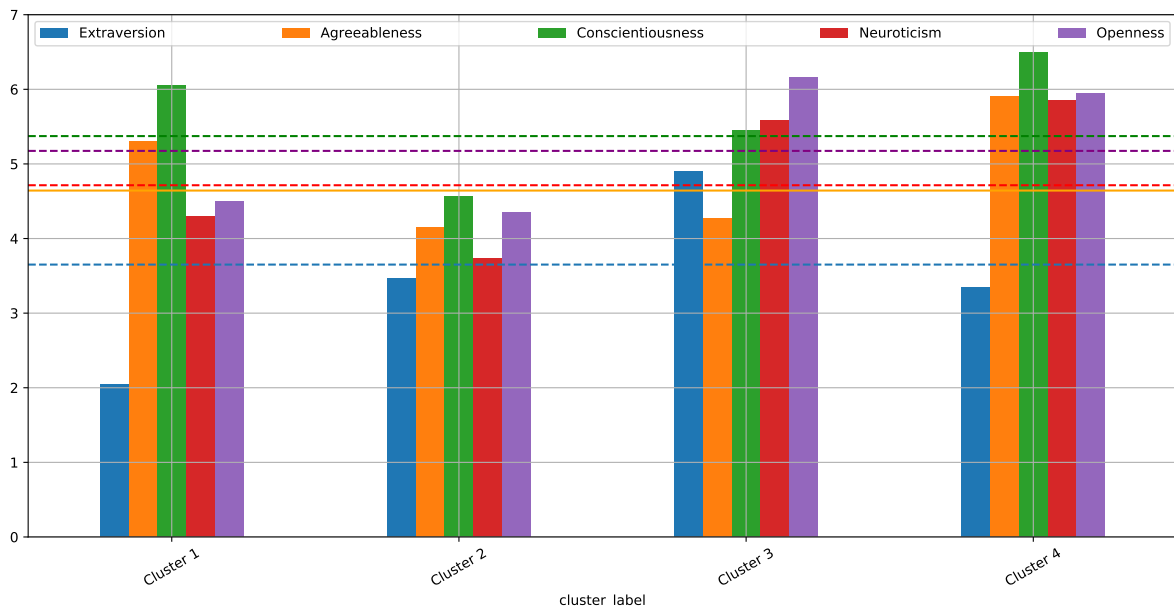


Figure 3.14: Average personality trait from the sample space and personality clusters.

Conscientiousness is on average the highest personality trait in the sample space, while *Extraversion* is least present. *Agreeableness* and *Neuroticism* are equally present, while *Open-*

ness is the second most present personality trait. *Cluster 1*, *Cluster 2*, and *Cluster 4* have *Conscientiousness* as the highest average personality trait. *Cluster 3* distinguish itself from the rest with having *Openness* as the highest personality traits. *Cluster 3* was also found to have higher *Extraversion* than other clusters.

3.3.2 User Evaluation

This section report the results of the analysis aimed at understanding the quality of recommending movies using facial expressions and emotions. The p-values were calculated for each question using a Proportional Two-Sided Test (2-sample z-test²) [30] to see if the proportions of selecting one recommendation approach over another is significantly different. This has been performed in addressing the research question:

RQ3: *In terms of Accuracy and Diversity, what is the quality of recommendation based on emotional responses, using facial expressions, in comparison to the other approaches?*

Accuracy

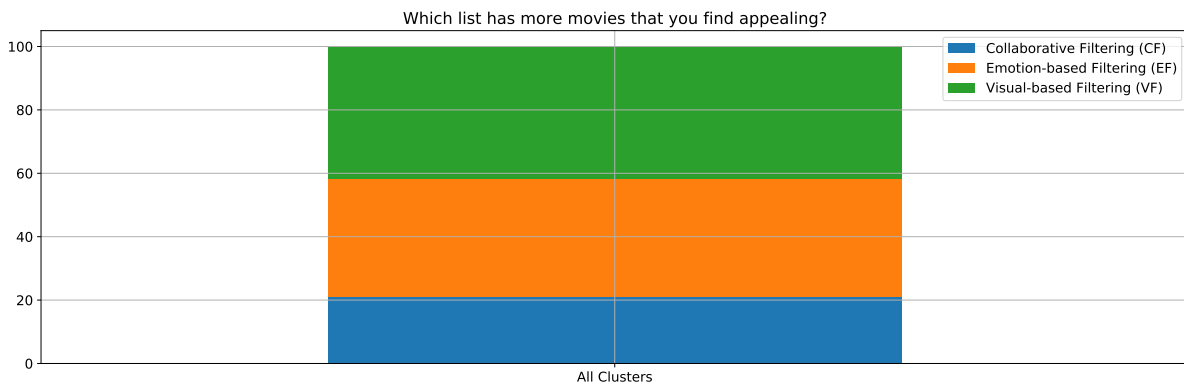


Figure 3.15: Selected Recommendation Technique by participants in generating appealing movie recommendations.

Table 3.3: Significance Evaluation: $H_0 = 0.05$

Comparison Condition (A vs. B)	Two-Sided Proportion Test (p-value)
Visual-based Filtering (VF) vs. Collaborative Filtering (CF)	0.037
Emotion-based Filtering (EF) vs. Visual-based Filtering (VF)	0.659
Emotion-based Filtering (EF) vs. Collaborative Filtering (CF)	0.096

Figure 3.15 and table 3.3 presents the results from asking the participants: *Which list has more movies that you find appealing?*

²https://www.statsmodels.org/stable/generated/statsmodels.stats.proportion.proportions_ztest.html

The majority of participants decided that Visual-based Filtering (VF) generated the most appealing movie recommendations with 42% of the votes. Emotion-based Filtering (EF) got 37% of the votes, while the baseline approach Collaborative Filtering (CF) got 21% of the votes.

From the calculated p-values, a significant difference in proportions between Visual-based Filtering (VF) and Collaborative Filtering (CF) was found. No significant difference between Visual-based Filtering (VF) and Emotion-based Filtering (EF) was found, while a marginal difference was found between Collaborative Filtering (CF) and Emotion-based Filtering (EF).

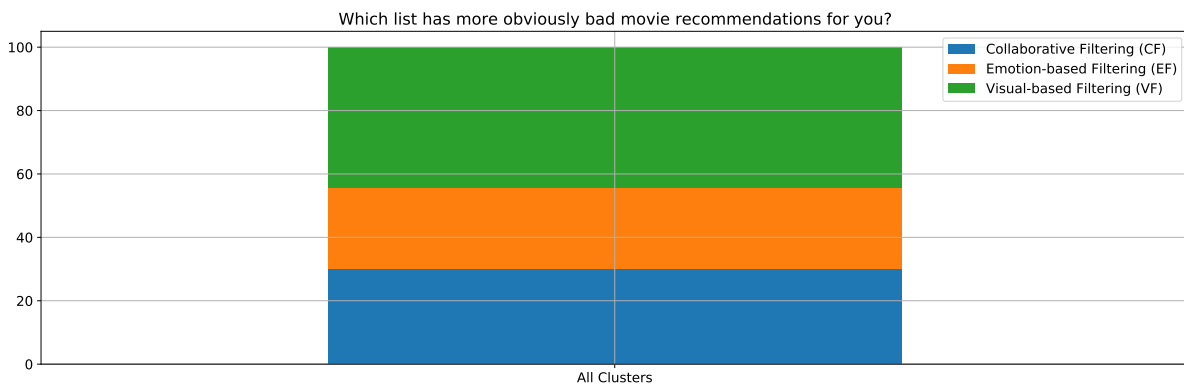


Figure 3.16: Selected Recommendation Technique by participants in generating bad movie recommendations.

Table 3.4: Significance Evaluation: $H_0 = 0.05$

Comparison Condition (A vs. B)	Two-Sided Proportion Test (p-value)
Visual-based Filtering (VF) vs. Collaborative Filtering (CF)	0.181
Emotion-based Filtering (EF) vs. Visual-based Filtering (VF)	0.070
Emotion-based Filtering (EF) vs. Collaborative Filtering (CF)	0.631

Figure 3.16 and table 3.4 presents the results from asking the participants: *Which list has more obviously bad movie recommendations for you?*

The majority of participants decided that Visual-based Filtering (VF) generated the most obviously bad movie recommendations with 44% of the votes. The baseline approach Collaborative Filtering (CF) got 30% of the votes, while Emotion-based Filtering (EF) got 26% of the votes.

From observing the p-values, no significant difference in proportions in regards to obviously bad movie recommendations were found. On the other hand, Visual-based Filtering (VF) and Emotion-based Filtering (EF) has a marginal difference in proportions.

When comparing the obtained results from the two questions, Visual-based Filtering (VF) was perceived to generate both appealing movies and bad movies. On the other hand, Emotion-based Filtering (EF) was found to generate appealing movies, and at the same time perceived by few of the participant to generate bad movie recommendations. With this observation, Emotion-based Filtering (EF) is voted to be perform best in terms of *Accuracy*.

Diversity

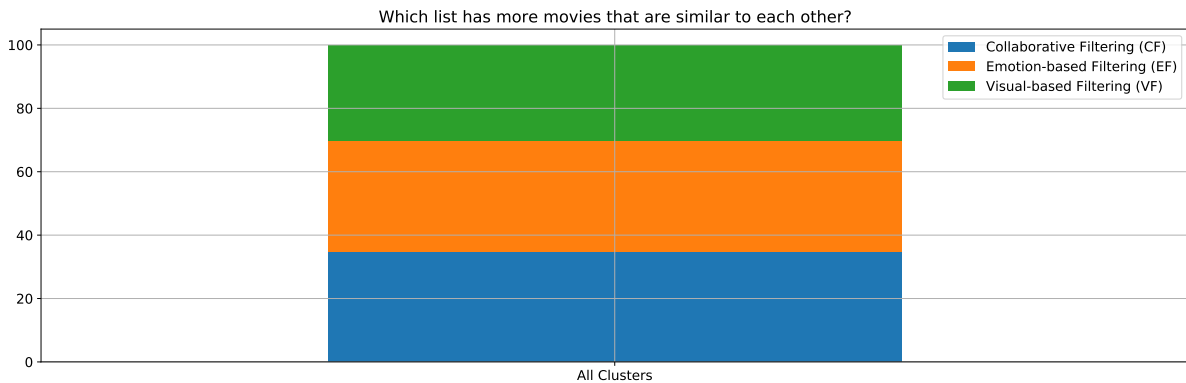


Figure 3.17: Selected Recommendation Technique by participants in generating similar movie recommendations.

Table 3.5: Significance Evaluation: $H_0 = 0.05$

Comparison Condition (A vs. B)	Two-Sided Proportion Test (p-value)
Visual-based Filtering (VF) vs. Collaborative Filtering (CF)	0.654
Emotion-based Filtering (EF) vs. Visual-based Filtering (VF)	0.654
Emotion-based Filtering (EF) vs. Collaborative Filtering (CF)	1.000

Figure 3.17 and table 3.5 presents the results from asking the participants: *Which list has more movies that are similar to each other?*

In this question, participants found both Collaborative Filtering (CF) and Emotion-based Filtering (EF) to recommend lists with similar movies in them with 35% of the votes each. Visual-based Filtering (VF) got slightly less with 30% of the votes.

The p-values calculated in regards to recommending movies that are similar to each other indicate no significant difference in proportions. Both Collaborative Filtering (CF) and Emotion-based Filtering (EF) share the same proportion, which makes the two other comparisons equally insignificant.

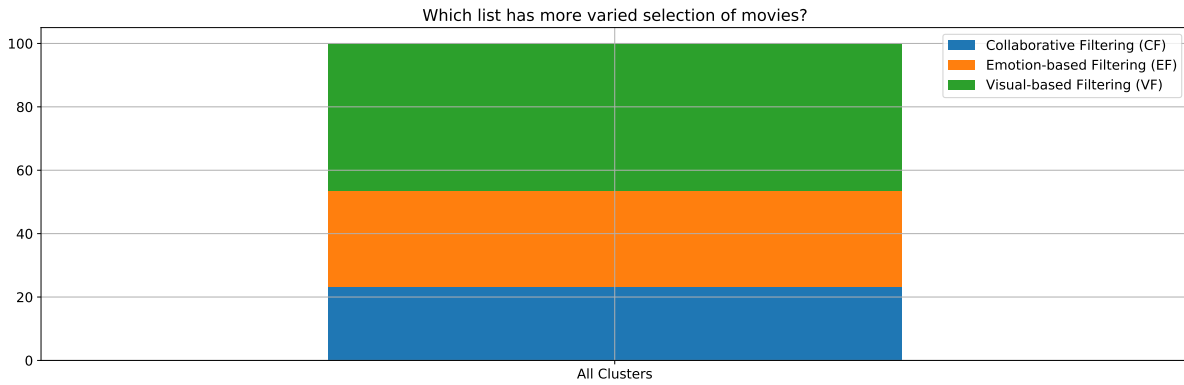


Figure 3.18: Selected Recommendation Technique by participants in generating varied movie recommendations.

Table 3.6: Significance Evaluation: $H_0 = 0.05$

Comparison Condition (A vs. B)	Two-Sided Proportion Test (p-value)
Visual-based Filtering (VF) vs. Collaborative Filtering (CF)	0.024
Emotion-based Filtering (EF) vs. Visual-based Filtering (VF)	0.121
Emotion-based Filtering (EF) vs. Collaborative Filtering (CF)	0.465

Figure 3.18 and table 3.6 presents the results from asking the participants: *Which list has more varied selection of movies?*

The majority of participants decided that Visual-based Filtering (VF) generated the most varied selection of movie recommendations with 47% of the votes. Emotion-based Filtering (EF) got 30% of the votes, while the baseline approach Collaborative Filtering (CF) got 23% of the votes.

From observing the calculated p-values, a significant difference in proportions between Visual-based Filtering (VF) and Collaborative Filtering (CF) was found, while no significant difference was found between Visual-based Filtering (VF) and Emotion-based Filtering (EF), and between Collaborative Filtering (CF) and Emotion-based Filtering (EF).

In terms of *Diversity*, the results obtained in this experiment show that Visual-based Filtering is significantly better than Emotion-based Filtering (EF) and Collaborative Filtering (CF). While Emotion-based Filtering (EF) is considered to have a bit more varied selection of movies compared to Collaborative Filtering (CF), it is too small of a difference to say anything conclusive. This makes them more or less equivalent in terms of *Diversity*.

3.3.3 User Evaluation by Personality

Going from the overall user evaluation of recommendation approaches, this section report the results of the analysis aimed at understanding how groups with similar personality traits evaluate the proposed recommendation approaches. This has been performed in addressing the research question:

RQ4: *In terms of Accuracy and Diversity, do users with similar personality traits prefer similar recommendation approaches?*

In the figures 3.19, 3.20, 3.21 and 3.22, we have stacked bars representing the distribution of selected answers. In addition to the bars, lines were added to represent the average personality traits for each group. When presenting the results for this section, we will observe if there is a majority of agreement on a preferred recommendation approach. Figure 3.23 shows a summary of the analysis, where the majority answers for each cluster is grouped into *Good* and *Bad* depending on the questions related to *Accuracy* and *Diversity*. However, the results does not indicate that groups with similar personality traits prefer similar recommendation techniques, except for cluster 4, which highly agreed on each evaluation question.

Accuracy

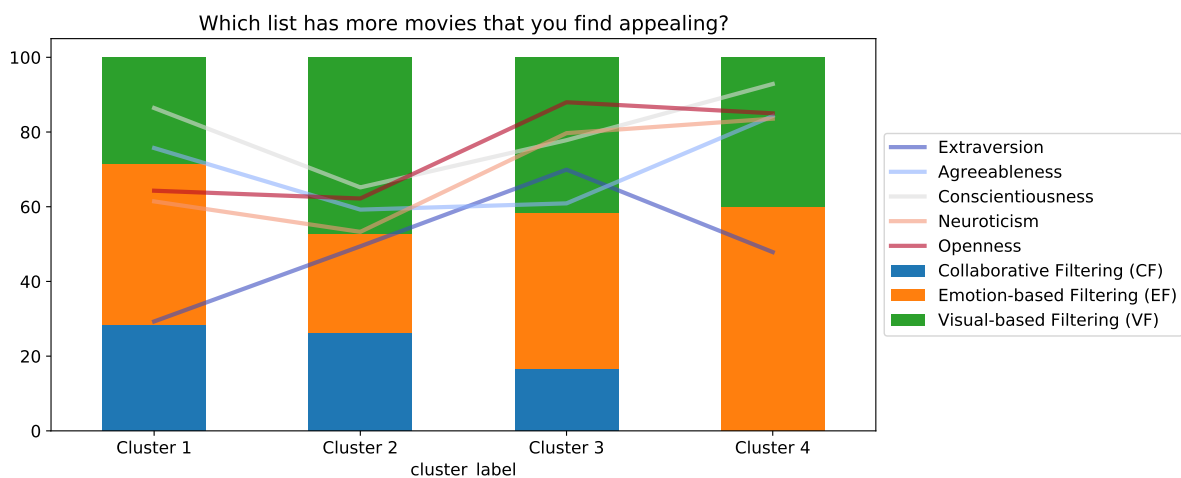


Figure 3.19: **Stacked bars:** Proportion of selected recommendation approach. **Lines:** Average personality.

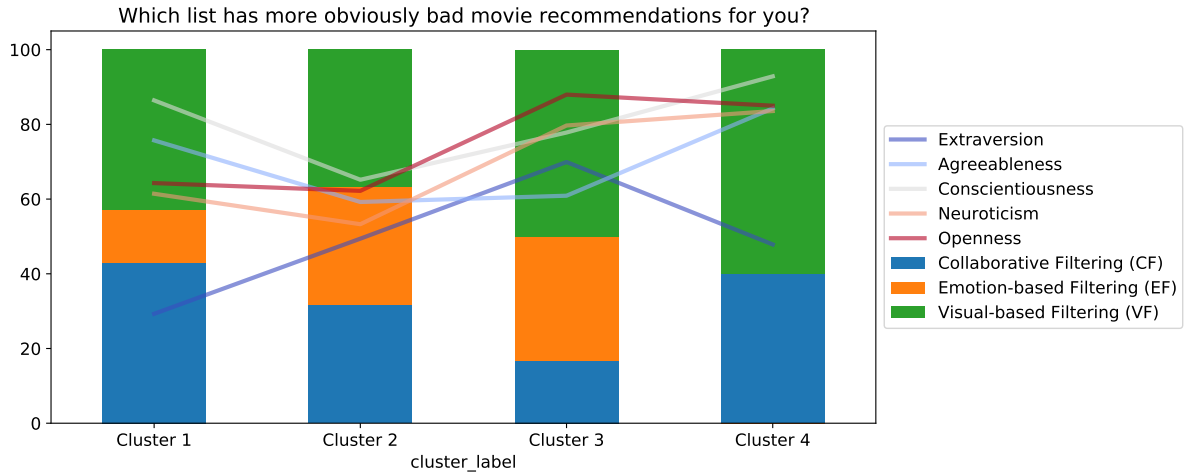


Figure 3.20: **Stacked bars:** Proportion of selected recommendation approach. **Lines:** Average personality.

Diversity

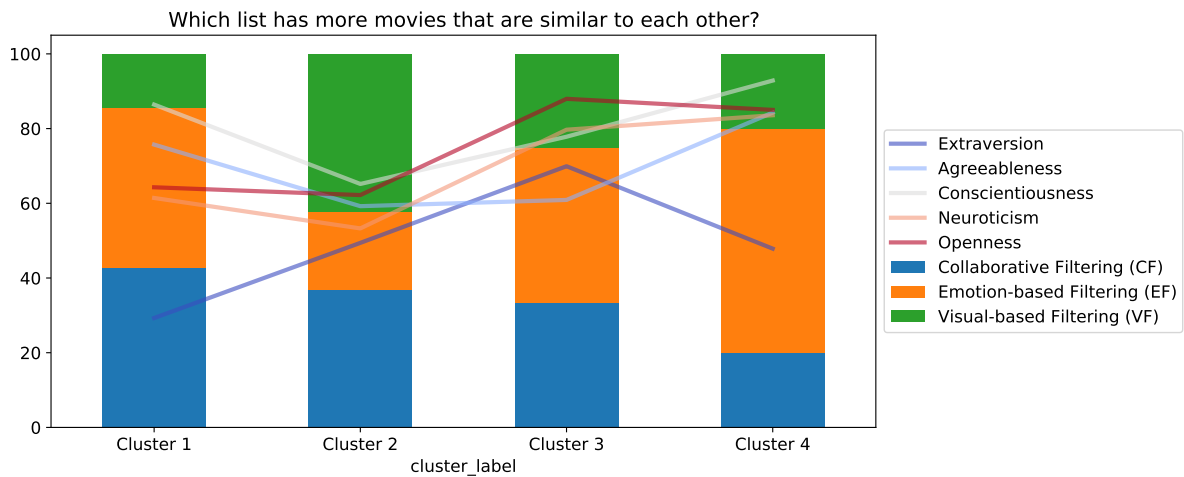


Figure 3.21: **Stacked bars:** Proportion of selected recommendation approach. **Lines:** Average personality.

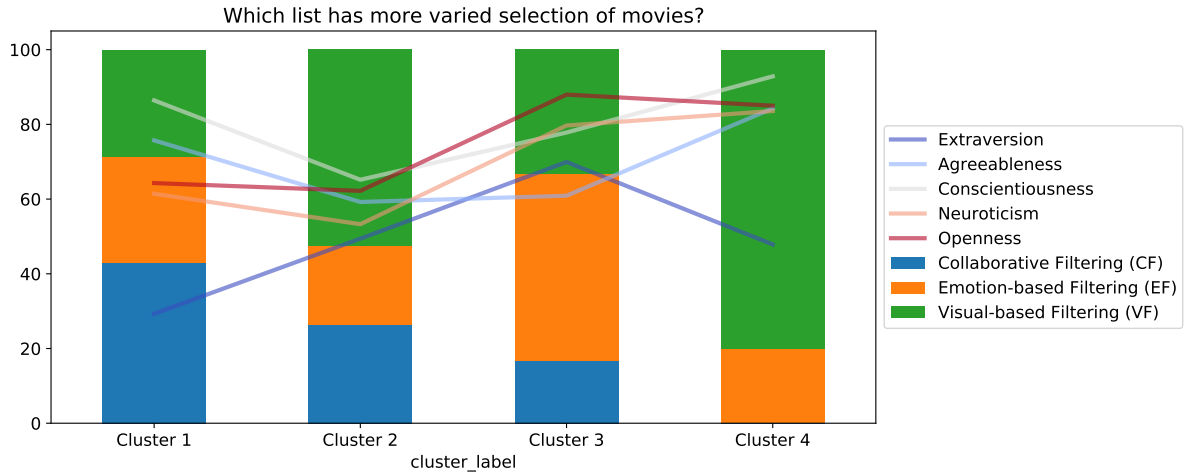


Figure 3.22: **Stacked bars:** Proportion of selected recommendation approach. **Lines:** Average personality.

	Accuracy		Diversity	
	Good	Bad	Good	Bad
Cluster 1	Emotion-based Filtering (EF)	Collaborative Filtering (CF) / Visual-based Filtering (VF)	Collaborative Filtering (CF)	Emotion-based Filtering (EF) / Collaborative Filtering (CF)
Cluster 2	Visual-based Filtering (VF)	Visual-based Filtering (VF)	Visual-based Filtering (VF)	Visual-based Filtering (VF)
Cluster 3	Emotion-based Filtering (EF) / Visual-based Filtering (VF)	Visual-based Filtering (VF)	Emotion-based Filtering (EF)	Emotion-based Filtering (EF)
Cluster 4	Emotion-based Filtering (EF)	Visual-based Filtering (VF)	Visual-based Filtering (VF)	Emotion-based Filtering (EF)
	Emotion-based Filtering (EF)		Visual-based Filtering (VF)	

Figure 3.23: Majority Selected Recommendation Approach. Green = Good performance, Red = Bad performance, Yellow = Indecisive Performance.

From observing the results in figure 3.23, the majority in Cluster 1 selected Emotion-based Filtering (EF) as having good *Accuracy*, while there was a split opinion about whether Collaborative Filtering (CF) or Visual-based Filtering (VF) was bad. Collaborative Filtering (CF) was also selected as good in terms of *Diversity*, but the cluster was indecisive in selecting which recommendation approach performed bad in terms of *Diversity*.

The majority of Cluster 2 selected Visual-based Filtering (VF) on all questions. The cluster seems to have weak consensus on which recommendation approach they preferred. Overall, the cluster seems to lean towards Visual-based Filtering (VF) as being good in both *Accuracy* and *Diversity* compared to the other recommendation approaches.

Cluster 3 was the cluster that distinguished themselves most in terms of personality traits compared to the other clusters. The cluster was indecisive in selecting a recommendation approach that had good *Accuracy*. On the other hand, we see that the majority selected Visual-based Filtering (VF) as performing bad in *Accuracy*, which indicate that the cluster tend to lean towards Emotion-based Filtering (EF) in having good *Accuracy*. For *Diversity*, Emotion-based Filtering (EF) was selected as good and bad.

Cluster 4 seems to have a higher consensus for each question compared to the other clusters. The cluster selected Emotion-based Filtering (EF) as good in terms of *Accuracy*, while Visual-based Filtering (VF) as bad. For *Diversity*, the cluster found Visual-based Filtering (VF) to be good, while Emotion-based Filtering (EF) was mostly perceived as bad.

The overall evaluation of the recommendation approaches was based on the majority selection, and the assumption of preferred approach from the negatively formulated questions. As an example, *Cluster 3* selected both Emotion-based Filtering (EF) and Visual-based Filtering (VF) as having good *Accuracy*, but also had a majority selecting Visual-based Filtering (VF) as recommending bad movies. By this, Emotion-based Filtering (EF) is assumed to be the preferred recommendation approach for *Cluster 3* in terms of *Accuracy*. With these assumptions, we found Emotion-based Filtering (EF) to have the best performance in terms of *Accuracy*, while Visual-based Filtering (VF) performed best in terms of *Diversity*.

3.4 Affective Preference Elicitation

In this section we report the results of the analysis aimed at understanding the feasibility in using facial expressions to elicit preferences and recommend movies. This has been performed in addressing the research question:

RQ5: *Can the preferences of users be elicited from their emotional responses extracted from the facial expressions in order to generate movie recommendations?*

This study set out to provide a proof of concept in using emotional responses and facial expressions of users to elicit preferences in the form of a rating. By using rigorously tested technologies, this research has constructed a method to collect facial expression while watching trailers, predict ratings using the facial expressions, and building affective profiles to be used in recommending movies. Before concluding the feasibility of using the method, we asked real users to participate in using the system.

The artifact developed was deployed and used by users online. The confirmation of having actual participants use the system, retrieving recommendations, and evaluating the proposed technique gave a proof of concept. When assessing the Emotion-based Filtering (EF),

the results from this experiment show that the technique performs better in terms of *Accuracy* than the baseline approach of Collaborative Filtering (CF).

To further test the feasibility in using observed emotions to estimate and elicit preferences, we retrained the models to estimate preferences with new data collected from the *main-study*. Instead of being constrained to a 5-Point rating scale, we experimented with a 3-Point rating scale, and *like*, or *dislike*. Every rating above or equal to 3 was considered a *like*, and every rating below 3 was considered *dislike*. Instead of 128 samples collected in the *pre-study*, we now had 406 samples of emotive features with a corresponding rating. Figure 3.7 shows the results obtained.

Model	Features	CV Accuracy Score (Avg)	CV Accuracy Score (Std)	Preference Scale
Random Forest Classifier	Max, Mean, Std	0.518	0.023	5-Point Rating Scale
Gradient Boosting Classifier	Min, Max, Std, Median	0.684	0.077	3-Point Rating Scale
Random Forest Classifier	Std, Median	0.890	0.035	Like, Dislike

Table 3.7: Best prediction score from retraining the models. Each row represents the results using different preference scales.

For the 5-Point rating scale, the score was almost the same as before, only *Random Forest Classifier* was found to be the best model. The average accuracy was 0.518 with a standard deviation of 0.023. Using a 3-Point rating scale, the accuracy went somewhat up, but with a higher standard deviation. With reducing the ratings to *like*, or *dislike*, *Random Forest Classifier* had an average accuracy of 0.89 with a standard deviation of 0.035. These results are promising, and show that emotional responses captured in the *consumption stage* is associated with the users explicit preferences. This means that preferences can be elicited from users emotional responses extracted from facial expressions in order to generate movie recommendations.

3.5 The System Usability

This section presents the obtained results from evaluating the usability of the artifact. In the last step of the experiment, participants were given a questionnaire to evaluate the system's usability. System Usability Survey (SUS) [4] was used for this evaluation. This survey presents 10 statements for the participants, which they select from a 5 or 7 point scale how much they agree with the statement. In this experiment, a 7 point scale was used as it is reported to

reflect a participant's opinion more precisely while not confusing the participant with too many options [19].

The score measured in SUS ranges from 0 to 100. A SUS score of 68 is considered as an *average* usability score, and the aim is to reach the score of 68 or above. By interpreting the score as percentiles, we can grade the overall usability of the system where a score of 68 is around a 50th percentile [5]. The *Curved Grading Scale for SUS* [27] is used in this study to grade the score from the System Usability Survey.

Grade	SUS	Percentile range
A+	84.1 - 100	96 - 100
A	80.8 - 84.0	90 - 95
A-	78.9 - 80.7	85 - 89
B+	77.2 - 78.8	80 - 84
B	74.1 - 77.1	70 - 79
B-	72.6 - 74.0	65 - 69
C+	71.1 - 72.5	60 - 64
C	65.0 - 71.0	41 - 59
C-	62.7 - 64.9	35 - 40
D	51.7 - 62.6	15 - 34
F	0 - 51.6	0 - 14

The final score was calculated using the same approach demonstrated by Brooke [4] with minor adjustments to the formula to make it correct using a 7 point scale. Instead of subtracting every even numbered statement by 5, the even numbered statements were subtracted by 7. The total score that was possible to get from a user was 60 points. The scores was than multiplied by 1.6666667 instead of 2.5 to fit in a scale from 0 to 100. This is the equivalent of the formula presented by Brooke [4] only adjusted to the 7 point scale [28].

The final score calculated was 66.78. The score is just below the recommended average of 68, and according to the grading scale give the usability of the system a grade of C. The score indicates that there is room for improvements in the usability of the system.

By the distribution of selected points in figure 3.24, we can see that there is a tendency to highly disagree if users would like to use the system frequently. Most participants find the system easy to learn, and find it easy to use. As the development of the system required technical solutions when working with web cameras and real-time affective computing, the result of participants finding the system easy to use without needing technical support or needing to learn many things, is a good achievement. Overall, there is room for improvement in usability, and the scores calculated from this research can be a good baseline for future improvements.

Statement	1. Disagree strongly	2. Disagree moderately	3. Disagree a little	4. Neither agree nor disagree	5. Agree a little	6. Agree moderately	7. Agree strongly
I think that I would need the support of a technical person to be able to use this system.	51%	27%	9%	4%	4%	0	4%
I thought there was too much inconsistency in this system.	9%	29%	18%	27%	9%	9%	0
I needed to learn a lot of things before I could get going with this system.	47%	27%	13%	0	7%	2%	4%
I found the system unnecessarily complex.	16%	16%	27%	18%	11%	4%	9%
I think that I would like to use this system frequently.	16%	20%	11%	20%	22%	9%	2%
I found the various functions in this system were well integrated.	0	2%	11%	29%	24%	27%	7%
I felt very confident using the system.	2%	4%	4%	13%	27%	36%	13%
I thought the system was easy to use.	0	4%	7%	7%	33%	33%	16%
I found the system very cumbersome to use.	13%	24%	9%	27%	20%	4%	2%
I would imagine that most people would learn to use this system very quickly.	0	7%	0	13%	18%	40%	22%

Figure 3.24: Distribution of answers from the System Usability Survey

Chapter 4

Conclusion

4.1 Summary

This master thesis has embarked on a challenging task that suggests using emotions captured through Affective Computing to estimating ratings. The Emotion-based Filtering (EF) technique was implemented and evaluated by real users. The goal was to develop a method which continuously obtained ratings without the intrusive aspect of Explicit Feedback, and to generate quality recommendations. The results show that using Emotion-based Filtering (EF) performed better in terms of *Accuracy* than the baselines. In terms of *Diversity*, the Visual-based Filtering (VF) technique was superior, while the baseline technique of Collaborative Filtering (CF) and the Emotion-based Filtering (EF) had equivalent performance.

When developing the Emotion-based Filtering technique (EF), we experimented with different Machine Learning models to predict ratings from facial expressions. While the prediction model developed in this project has shown to work adequately, there is still more to be done. Some of the challenges with training a reliable prediction model revolve around obtaining enough data to generalize the estimations. The *pre-study* of this research conducted a small data collection phase to obtain such data. However, in obtaining these data, one may argue that the amount obtained was too small to generalize the prediction accuracy. Having said that, the overall performance of the model developed in this research indicates promising results in both offline evaluations and online evaluations.

In evaluating the recommendation techniques, previously used surveys-questions to evaluate recommendation algorithms were adopted. Users compared and evaluated three recommendation lists generated by the recommendation techniques. The evaluation was also extended to find relationships between personality similarities in users and the choice of recommendation technique.

In addition, this research aimed at finding relationships between visual features of movies, and the emotional responses of users. The objective demanded processing and structuring of two datasets so that the sequence of visual features in movie trailers matched with the sequence of emotional responses. The procedure to achieve this objective was time-consuming and required research in Correlational Methodology and Metadata Analysis. The results found from completing the objectives of this project has created several aspect to discuss.

4.2 Discussion

This experiment has demonstrated that facial expressions and emotions can be used as contextual information for Implicit Feedback. While eliciting preferences in this manner differs from Explicit Feedback, the technique can substantially reduce limitations of previous preference-elicitation techniques. While we can not conclude that the preferences obtained have the same quality as when using Explicit Feedback, the effect of this experiment show that ratings can be obtained unobtrusively as a temporarily solution until users explicitly provide their preferences. This will likely enhance the experience of using a recommendation system, as it will be less prone to repeatedly recommend old, highly rated content. The recommendation system can quickly start acquiring preferences in the background, which will reduce the time to overcome the cold-start limitation and data sparsity.

In many ways, estimating preferences based on emotions only tries to understand how the continuous change in emotions associate with ratings. Which means, we can not compare these types of estimated ratings with the actual ratings a user provide. If the appearance of smile and expressed joy are found to be associated with ratings of 5, this does not necessarily mean that a user would rate it 5. It is only a probabilistic assumption to help the recommendation engine. Hence, recommendations generated by such preferences should be explained to users.

In comparison to other techniques that uses contextual information besides the user preferences, the use of Emotion-based filtering has its roots in the content being consumed. Hence, similarities in users and items are found through inferred preferences, and not any contextual information not related to the content. This is likely to improve accuracy in finding similarities among users and items, and generate more accurate recommendations.

The results from Emotion-based filtering brings intriguing applications into other domains. Ratings are used to describe preferences in multiple domains. Hence, observing facial expressions and emotions when reading articles, listening to music, looking at aesthetic objects such as art and decorate items can be exploited to understand preferences of people.

These applications of utilizing emotions can help users to find emotionally-based recommendations. While these ambitions are strong, information systems understanding users emotions can be a big leap in decision making and personalized recommendation. In E-commerce, users can better evaluate decisions based on the perceived observation of emotion by the system, and their actual opinion.

In addition to the promising applications of using Emotions-based filtering, this research contributes to further analysis of emotional responses and the relationships with visual features. In the domain of movies, joy was found to be more prominent in more positive genres (i.e., Comedy, Children, Romance), while disgust was more prominent in negative genres (i.e., Horror, War, Action). These emotions can be indicative to understand a users affective state in the *entry stage*, the *consumption stage*, and the *exit stage* explained by Tkalčič et al. [42]. How to use these findings are still in its infancy and needs further study, but they establish grounds to work from. We still need to understand if observed emotions in users encourage recommendations of similar emotionally-induces content. The order of transitions between emotional states might have an impact on what to recommend. If joy is the observed emotions, we need to understand if users want more of joy-induced content, or if other emotionally-induced contents are appropriate to recommend.

Correlations between the visual features of movies and the emotional responses of users can improve recommendations by filtering movies or genres that has a tendency to induce certain emotions. If a user often likes horror movies, and is observed to express the emotion of surprise, we can filter horror movies measured to have high saturation variation. That is, the visual feature found to have a correlation with an emotional response within a genre. More specifically, this filtering can be used on new movies, to find users which like similar movies based on their emotional correlations to movies with similar visual features. Basically, to filter movies similar to a users preferences in visual features, and the expected emotional responses. This can possibly help the recommendation system to estimate the likelihood of liking or disliking a new item.

In studying the evaluation based on participants personality traits, clusters tended to agree in varying degree when evaluating a recommendation technique. In comparing all the cluster, *Cluster 4* was found to have the strongest agreement, while the other clusters were more ambiguous in their evaluation. With these findings, there is potential to customize recommendation techniques based on users personality traits. However, the study also suggest it to be mostly applicable for clusters which show clear indications of a preferred recommendation approach. While the results did find differences, the overall analysis suggests that Emotion-based Filtering (EF) performed better in terms of accuracy, while Visual-based Filtering (VF) performed better in terms of diversity.

With the presented implementation of an Emotion-based filtering technique which continuously monitor users facial expressions and emotions when consuming content, this research can hopefully be used as a guideline on how to implement such an approach to other projects. The capturing of emotions by this approach brings with it some benefits. Affective user profiles will continuously get larger as users engage with the system. This leads to more data and preferences which can be studied. The model for predicting ratings can be retrained as more data is available, and possible generalize better. The prediction model can also be trained on each users affective profile as it gets bigger. Eventually, affective preference profiles can be compared among users to find similar users. Lastly, the approach require little effort from users, and can ultimately be utilized as a great asset in filtering relevant content. While there are benefits in the proposed approach, we still don't know how users feel about being observed in this manner. Users will likely want the option to opt out of such techniques when not comfortable with being observed. Even though the observations are only data points, and not any real recording of their face, these types of issues needs to be further addressed and studied.

4.3 Future Work

For future work, directions have been divided into short-term, medium-term, and long-term. In the short term, the project should be extended with more participants to see if there is a change in how the recommendation techniques are evaluated.

Furthermore, with more participants, new opportunities to fit the recommendation approach for each user occurs. For example, in the medium term, we can investigate if training a prediction model for each user's affective profile will improve the quality of estimating ratings. This is based on the assumption that we all express our emotions uniquely. This also implies that participants will likely need to watch more than minimum of four trailers to build stronger affective profiles. With running the experiment multiple times over a time-period, we can compare results and see if participants find the recommendations to improve as their affective profiles gets more ratings. If this is the case, then the method of implicitly obtaining ratings through facial expressions can reduce the affect of cold-start for new users.

In the long-term, the MovieLens dataset should be gradually replaced with a dataset of only emotionally estimated ratings. This is so that recommendations algorithms can filter users with similar preferences solely on their affective preference profiles. In order to achieve this, we want to experiment with reducing the available movies within the system to obtain more overlap of facial expressions and estimated ratings. Utilizing crowd-sourcing can also contribute in building larger datasets. Eventually, we aim to explore the power of using visual

features combined with emotional responses as descriptive features to generate quality recommendations.

4.4 Acknowledgement

"This work was supported by industry partners and the Research Council of Norway with funding to MediaFutures: Research Centre for Responsible Media Technology and Innovation, through The Centres for Research-based Innovation scheme, project number 309339."

References

- [1] Affectiva (2017, October). Emotion AI 101: All About Emotion Detection and Affectiva's Emotion Metrics. <https://blog.affectiva.com/emotion-ai-101-all-about-emotion-detection-and-affectivas-emotion-metrics>. Accessed: 2021-05-22.
- [2] Aggarwal, C. C. (2016). An introduction to recommender systems. In *Recommender systems*, pp. 1–28. Springer.
- [3] Barjasteh, I., R. Forsati, F. Masrour, A.-H. Esfahanian, and H. Radha (2015). Cold-start item and user recommendation with decoupled completion and transduction. In *Proceedings of the 9th ACM Conference on Recommender Systems*, pp. 91–98.
- [4] Brooke, J. (1996). Sus: a “quick and dirty” usability. *Usability evaluation in industry* 189.
- [5] Brooke, J. (2013). Sus: a retrospective. *Journal of usability studies* 8(2), 29–40.
- [6] Burke, R. (2002). Hybrid recommender systems: Survey and experiments. *User modeling and user-adapted interaction* 12(4), 331–370.
- [7] Deldjoo, Y., M. Elahi, P. Cremonesi, F. Garzotto, P. Piazzolla, and M. Quadrona (2016). Content-based video recommendation system based on stylistic visual features. *Journal on Data Semantics* 5(2), 99–113.
- [8] Deldjoo, Y., M. Elahi, M. Quadrona, and P. Cremonesi (2018). Using visual features based on mpeg-7 and deep learning for movie recommendation. *International journal of multimedia information retrieval* 7(4), 207–219.
- [9] Deldjoo, Y., M. Schedl, and M. Elahi (2019). Movie genome recommender: A novel recommender system based on multimedia content. In *2019 International Conference on Content-Based Multimedia Indexing (CBMI)*, pp. 1–4. IEEE.
- [10] Dunlap, W. P., M. B. Jones, and A. C. Bittner (1983). Average correlations vs. correlated averages. *Bulletin of the Psychonomic Society* 21(3), 213–216.

- [11] Elahi, F. B. M. M. et al. (2019). Cold start solutions for recommendation systems.
- [12] Elahi, M. (2014). *Empirical evaluation of active learning strategies in collaborative filtering*. Ph. D. thesis, Ph. D. Dissertation. Ph. D. Dissertation. Free University of Bozen-Bolzano.
- [13] Elahi, M. (2019). Cold start solutions for recommendation systems music recommender systems view project extra: Expertise-boosted model for trust-based recommendation system based on supervised random walk view project.
- [14] Elahi, M., F. Bakhshandegan Moghaddam, R. Hosseini, C. Trattner, and M. Tkalčič (2019, 06). Ma14kd [original] dataset description: Visual attraction of movie trailers.
- [15] Elahi, M., M. Braunhofer, T. Gurbanov, and F. Ricci (2018). User preference elicitation, rating sparsity and cold start.
- [16] Elahi, M., M. Braunhofer, F. Ricci, and M. Tkalčič (2013). Personality-based active learning for collaborative filtering recommender systems. In *Congress of the Italian Association for Artificial Intelligence*, pp. 360–371. Springer.
- [17] Elahi, M., F. Ricci, and N. Rubens (2014). Active learning strategies for rating elicitation in collaborative filtering: A system-wide perspective. *ACM Transactions on Intelligent Systems and Technology (TIST)* 5(1), 1–33.
- [18] Fernández-Tobías, I., M. Braunhofer, M. Elahi, F. Ricci, and I. Cantador (2016). Alleviating the new user problem in collaborative filtering by exploiting personality information. *User Modeling and User-Adapted Interaction* 26(2), 221–255.
- [19] Finstad, K. (2010). Response interpolation and scale sensitivity: Evidence against 5-point scales. *Journal of usability studies* 5(3), 104–110.
- [20] Fletcher, K. K. (2017). A method for dealing with data sparsity and cold-start limitations in service recommendation using personalized preferences. In *2017 IEEE international conference on cognitive computing (ICCC)*, pp. 72–79. IEEE.
- [21] Gosling, S. D., P. J. Rentfrow, and W. B. Swann Jr (2003). A very brief measure of the big-five personality domains. *Journal of Research in personality* 37(6), 504–528.
- [22] GroupLens (2003, February). MovieLens 1M Dataset. <https://grouplens.org/datasets/movielens/1m/>. Accessed: 2021-05-22.
- [23] Hevner, A. R., S. T. March, J. Park, and S. Ram (2004). Design science in information systems research. *MIS quarterly*, 75–105.

- [24] Isinkaye, F. O., Y. Folajimi, and B. A. Ojokoh (2015). Recommendation systems: Principles, methods and evaluation. *Egyptian informatics journal* 16(3), 261–273.
- [25] Joho, H., J. M. Jose, R. Valenti, and N. Sebe (2009). Exploiting facial expressions for affective video summarisation. In *Proceedings of the ACM international conference on image and video retrieval*, pp. 1–8.
- [26] Kim, H.-N., A. El-Saddik, and G.-S. Jo (2011). Collaborative error-reflected models for cold-start recommender systems. *Decision Support Systems* 51(3), 519–531.
- [27] Lewis, J. R. and J. Sauro (2018). Item benchmarks for the system usability scale. *Journal of Usability Studies* 13(3).
- [28] Lewis, J. R. and J. Sauro (2020). Converting rating scales to 0–100 points. <https://measuringu.com/converting-scales-to-100-points/>. Accessed: 2021-03-25.
- [29] Lika, B., K. Kolomvatsos, and S. Hadjiefthymiades (2014). Facing the cold start problem in recommender systems. *Expert Systems with Applications* 41(4), 2065–2073.
- [30] Ludbrook, J. (2013). Should we use one-sided or two-sided p values in tests of significance? *Clinical and Experimental Pharmacology and Physiology* 40(6), 357–361.
- [31] Ma, C.-C. (2008). A guide to singular value decomposition for collaborative filtering. *Computer (Long Beach, CA) 2008*, 1–14.
- [32] Moghaddam, F. B., M. Elahi, R. Hosseini, C. Trattner, and M. Tkalčić (2019). Predicting movie popularity and ratings with visual features. In *2019 14th International Workshop on Semantic and Social Media Adaptation and Personalization (SMAP)*, pp. 1–6. IEEE.
- [33] Nielsen, J. (1994, April). 10 Heuristics for User Interface Design: Article by Jakob Nielsen. <https://www.nngroup.com/articles/ten-usability-heuristics/>. Accessed: 2021-05-22.
- [34] Rashid, A. M., I. Albert, D. Cosley, S. K. Lam, S. M. McNee, J. A. Konstan, and J. Riedl (2002). Getting to know you: learning new user preferences in recommender systems. In *Proceedings of the 7th international conference on Intelligent user interfaces*, pp. 127–134.
- [35] Rimaz, M. H., R. Hosseini, M. Elahi, and F. B. Moghaddam. Audiolens: Audio-aware video recommendation for mitigating new item problem.
- [36] Rubens, N., M. Elahi, M. Sugiyama, and D. Kaplan (2015). Active learning in recommender systems. In *Recommender systems handbook*, pp. 809–846. Springer.

- [37] Schein, A. I., A. Popescul, L. H. Ungar, and D. M. Pennock (2002). Methods and metrics for cold-start recommendations. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 253–260.
- [38] Sidorov, G., A. Gelbukh, H. Gómez-Adorno, and D. Pinto (2014). Soft similarity and soft cosine measure: Similarity of features in vector space model. *Computación y Sistemas* 18(3), 491–504.
- [39] Silver, N. C. and W. P. Dunlap (1987). Averaging correlation coefficients: should fisher's z transformation be used? *Journal of applied psychology* 72(1), 146.
- [40] Tkalčič, M., M. Elahi, N. Maleki, F. Ricci, M. Pesek, and M. Marolt (2019, 3). Prediction of music pairwise preferences from facial expressions. In *International Conference on Intelligent User Interfaces, Proceedings IUI*, Volume Part F147615, New York, NY, USA, pp. 150–159. Association for Computing Machinery.
- [41] Tkalčič, M., N. Maleki, M. Pesek, M. Elahi, F. Ricci, and M. Marolt (2017). A research tool for user preferences elicitation with facial expressions. In *Proceedings of the eleventh acm conference on recommender systems*, pp. 353–354.
- [42] Tkalčič, M., A. Kosir, and J. Tasic (2011). Affective recommender systems: the role of emotions in recommender systems. In *Proc. The RecSys 2011 Workshop on Human Decision Making in Recommender Systems*, pp. 9–13. Citeseer.
- [43] Warriner, A. B., V. Kuperman, and M. Brysbaert (2013). Norms of valence, arousal, and dominance for 13,915 english lemmas. *Behavior research methods* 45(4), 1191–1207.
- [44] Zhang, S., W. Wang, J. Ford, F. Makedon, and J. Pearlman (2005). Using singular value decomposition approximation for collaborative filtering. In *Seventh IEEE International Conference on E-Commerce Technology (CEC'05)*, pp. 257–264. IEEE.