Using Biometric Signals and Virtual Reality to Evaluate Movies and/or TV Shows

Daniel Filipe Garcia Gonçalves
daniel.f.g.goncalves@tecnico.ulisboa.pt

Instituto Superior Técnico, Lisboa, Portugal

October 2022

Abstract

Every multimedia asset has one major goal: Invoke an emotional reaction on the viewer. Furthermore, they are even judged for how well they perform this task. Movies and TV Shows are some of the major types of multimedia content that are fully structured and developed to attend to the viewers' emotional needs or to enter their personal world by promoting some kind of psychological influence. With the growth of Virtual Reality, several studies have been made about what the future of this area will be in regards to the movie industry and how it is expected to become the main approach to watching movies, even when compared to the cinema. When trying to evaluate the impact of a movie next to the audience, the industry tends to use questionnaires or other types of subjective information gathering methods, resulting in ambiguous feedback and wrong content classification. In this thesis, a new approach is presented with the intent to to use biometric signals taken directly from the viewers, at the precise moment of watching some video content, and use it to measure the viewer's arousal during the whole watching experience, providing a new data point to be used for multimedia artifact classification. Virtual Reality is also a decisive factor, since it will guarantee greater presence, less distraction induced measuring errors and a glimpse into the future of the industry.
Keywords: Movie; TV Show; Arousal; Virtual Reality; Biometric Classification; Emotion Recognition.

1. Introduction

With the growth of Movie/TV Show streaming, audiences have become the primary object of study when evaluating the effectiveness or quality of such multimedia content. As Marine Boulanger, Médiamétrie's Director of Cinema, pointed out [10]: *"The cinema is a place for emotions. Our research shows that among some 42 million cinemagoers in France in 2018, 3 out of 4 viewed the cinema primarily as a catalyst for emotions."*

Although the movie industry has suffered from a boom of new technologies and ways to develop itself, such as new watching devices, ways to distribute content, or even real-time personalized storytelling methods, the impact of the provided content next to the audience is still a tricky factor to measure. Let's look at the contemporary techniques of formally requesting feedback from the audience regarding some Movies or Tv Shows. The main ones are Focus groups, Critics' reviews, Online Reviews, Anonymous Questionnaires, and user data retrieved by the streaming platforms, such as watch-time or simple qualitative evaluations. These methods have one thing in common: they all return conscious results. Results such as these are deeply influenced by many factors that may make them useless or even invalid, bringing to the surface important questions about the excess of Ambiguity, lack of Precision, and Accuracy.

By increasing the accuracy of the contemporary feedback retrieval methods, the industry would not only provide a better experience to the viewer, but it would also contribute to predicting the audience response to an artifact better and, consequently, lead to better executive decisions, acceptable tuned creative paths, and well-supported marketing strategies.

2. Background

2.1. The Movie Industry

Before COVID-19, the movie industry was in what is known as its *historic period*, between 2015 to 2020. During this period, the movie industry reached about $234.9 billion in value and has achieved a compound annual growth rate of 2.4%, presenting a predicted market value of $318.5 billion and $410.6 billion by 2025 and 2030, respectively Experts have even indicated that the video streaming market should reach $223.98 billion by 2028, with a CAGR of 21.0% [5]. This pronounced

growth in just five years has been justified by several social, technological, and even demographic factors [6].

With the growth of streaming, movies and TV shows are suddenly closer than ever to the public, providing a much more personalized experience by giving control to the viewer, not of the content itself, but of How and When it is watched. The How part is exciting, considering that this new proximity to the viewers seems to have brought a lot of unique content-watching methods to the table.

2.2. Model of Emotions

Since the proposed solution intensely focuses on the impact a multimedia artifact generates on a viewer, it's essential to evaluate how such an effect may be assessed. Psychology and Effective Science have been trying to find a descriptive yet visual way to define human emotion. Although there seem to be a lot of different proposals, there are two main areas in which all of them seem to be included: The Discrete Emotions Theory and the Dimensional Theory of Emotion.

The first indicates that emotions can be characterized in finite designations. In contrast, the second one refuses this idea and proposes that a dimensional model should be used to better account for slight variations and interdependence between emotional states.

Between the already exposed Dimensional Theories of Emotion, there is one that seems to be the most referenced, which was presented by Russel [11] and uses two different axis: Arousal, which represents the vertical axis and is commonly associated with the intensity of an emotion, and Valence, which represents the horizontal axis.

This Arousal feature is one of the primary resources of the system developed and presented in this document.

2.3. Virtual Reality

If we look at the Conviva Q4 2020 report [7], we can see that TVs only take about 17% of the watching time, globally, and the other 83% is distributed throughout other technological mediums. This shows that a conclusion can be taken: Streaming opens doors for new and innovative technologies that can improve the watching capabilities of some movies or TV shows and even dialogue with such content to provide a more immersive and accurate to the initial purposes' viewing experience.

One of these technologies is Virtual Reality. When we talk about the use of Virtual Reality in the movie industry, we are automatically talking about the future, but maybe not a very distant one. Examples of the use of Virtual Reality to improve this industry and create new applications, consequently, expanding it, are plenty and widely available with a single search online. Watching traditional movies with an HMD and a specialized and personalized video player like SKYBOX [2], or simply entering the world of a 360^{o} movie through YouTube, are just some examples of already widely used connections between Virtual Reality and Movies/TV Shows. Connections between this technology and streaming are also present.

2.4. Virtual Reality as a movie/TV show generated emotion amplifier

Two central psychology-related concepts directly connect to Virtual Reality: Immersion and Presence.

Slater and Wilbur [13] have pointed out that the correlation of four factors can define immersion: inclusion, extension, surround effect, and vividness. As stated by Slater [12], Presence is also related to the illusion of being transported into the new virtual world. In contrast with immersion, it is a perceptual concept, not a technological/technical one.

We can conclude that presence is directly related to the viewer's reaction, while immersion is a much more Virtual Reality experience-related metric.

Suppose we try to find a relation between the emotions provoked by a movie and these two Virtual Reality related metrics. In that case, we can quickly conclude that a more considerable immersion can lead to a more significant presence and a more prominent presence directly leads to more robust and genuine emotions. This proportional correlation between immersion and the emotional reactions performed by the viewers was already tested and proved in a report by T Vish., S. Tan, and Molenaar [14].

2.5. Movie/TV Show classification

We've already seen that a movie or TV Show is intended to generate a sequence of emotional states next to the viewer, that immersion and presence do increase the power of such emotions, and that the ones caused by perception are the truest and valid due to their unconscious source so, when we think about how the audience can classify such multimedia artifacts, we should find out that such classification would be made based on perceptual responses, but that's not what happens. All of them seem to be a result of conscious thinking. The main mediums by which a viewer is asked to classify some movie or TV show are the following:

- **Focus Groups**: In regards to the results portrayed by this method of feedback gathering, the possibly damaging factors start even before the watching experience: The person may be having a bad day, and this will make them have more tendency to lower arousal values/negative emotions since these screenings are a previ-

ously scheduled event and do not care about the individually will that the viewers show at that exact moment; the viewer may be provided with few to no information about what they are watching, what can result in stress, anxiety and this may affect their comments after the movie; an NDA (non-disclosure agreement) can also be asked to the viewer, resulting in significant amounts of stress and legal pressure and force them to not answer truthfully to the questions made to them at the end; the whole formal and exclusive environment, both social and physical, can also influence the viewer's mood and, consequently, the feedback given by the viewer. Another significant factor is related to the choosing of a representative audience for a test screening. Companies tend to choose a diverse group of people to achieve better results, which is a good practice. Still, the number of participants will never be significant since a few hundred chosen viewers do not represent the billions that exist. Another big problem is related to the content shown because it is usually not the final version, which can seriously impact the provided feedback.

- **Critic's/Media Outlet's Reviews**: This type of specialist-provided reviews can also lead the audience to incorrectly classify a movie and spread such an idea without even watching it. This would be a little less bad if the critic's opinions did not diverge from the audience; it does not seem to be the case. Film Data Researcher Stephen Follows pointed out in an article named "Are film critics losing sync with audiences?" [4] that, in fact, critics' opinions have been gradually diverging from the public's evaluations of the same artifact. This article also shows another concerning fact: The genres that appease audiences are very different from those that the critics prefer, and differences such as this seem to aggravate through the years. Apart from this, we keep having the audience representation issue, as critics are in far less quantity than the actual possible viewers of some multimedia artifact. Regarding Media Outlet's reviews, not only do they still have the same problems stated before, but some questions about independence and impartiality should also be considered, since several movie companies also own major media outlets or participate in the movie review industry.
- **Online Reviews**: This method is the one that, by giving the most freedom for the general audience to criticize some content, suffers from the drawbacks of such a high degree of freedom. Viewers that, apart from enjoying or not some movie, provide a correct argument review have their thoughts mixed up with Fans that would do everything to exacerbate the movie's characteristics and haters that constantly try to put down some movie, TV Show, or franchise, director, actor, etc. Although, there is some silver lining in this approach because this high level of freedom leads to an increased audience representation, at least in number, by a lot compared to the previously presented methods of feedback retrieval.
- **Streaming Recommendation Systems**: As stated in the literature [9], there are two main types of Movie/TV Show recommendation systems: Content-Based, in which similarities between multimedia artifacts (director, actors, genre, etc.) classify them as suitable for a single viewer or not, and Collaborative Filtering, in which the habits and preferences of every viewer influences the choice of a single artifact for a single viewer. Both systems use simple objective parameters to classify content, such as watching time, genre, etc., but these parameters can also generate wrong results. An example: A viewer starts to watch an episode of some TV Show, but during the experience, receives a call and has to turn off the TV. Suddenly, the algorithm stores this as an incomplete watching period, which leads to classifying that TV Show as probably not suited for them. This simple occasion will also contribute to not recommending that content to other viewers. This collaborative contribution may be small, but it exists. With this example, we can understand that external factors resulting from low immersion values can influence the classification process in a streaming environment. Other subjective metrics are used for ranking and classifying multimedia artifacts, such as the traditional "like button," which can also introduce questions like personal understanding.
- **Viewer Questionnaires**: This is the most used method for viewer feedback gathering. These questionnaires can vary in application, type of questionnaire, and even in being mandatory or not. Starting with the first factor, a questionnaire can be delivered in almost every possible situation: After a focus group test screening, as an add-on to an online review, or randomly after a user of some streaming platform watches an episode of a TV show. The two main problems with such questionnaires seem to be, once again, several answers since you can't possibly have every viewer of some movie or TV show answer a questionnaire, and we also have a significant problem: Personal interpretation, and not just by the

viewers, but by the author as well. When someone is writing a questionnaire, passing thoughts or objective metrics to words and structuring those words into accessible phrases can result in imprecise questions. The viewer can then misunderstand these questions, which results in incorrect responses. Another problem can be quantitative measures asked the viewers since a simple scale of numbers is not enough to account for every possible understanding of the question's wording. Another psychological factor is the viewer's mood when answering the questionnaire, influencing the given answers.

3. Proposed System

The setup presented in this section intends to classify a particular movie clip with a continuous value from one to seven, related to the arousal felt during the process, based on the analysis of bio-signals extracted from a set of individuals while watching such a clip. To gather Electrocardiogram-related data to be used when calculating some Heart Rate Variability-related metrics, twenty health participants were chosen to interact with the experimental setup. Each participant watched two sets of clips, one in a traditional way and another in a Virtual Reality environment. This data was then used to train a Machine Learning algorithm, more precisely, a sequential model, utilizing the arousal values presented by an already developed study [8], since the used movie clips also came from the same study. In another phase of testing, and to prove that arousal-related classifiers would be helpful for the users of movie/TV show streaming apps, a prototype of such type of app was developed and exposed to a usability test with eight participants, and not the same ones from the first experimental test.

3.1. Data Gathering Setup

To record the Electrocardiogram data from each individual, an hardware setup was developed comprised of an Arduino Uno board, an AD822 module, and an HC-06 module. This setup allowed for data gathering and wireless sending to the experimental laptop's Heart Rate Variability python algorithm in real-time. Both sensor placement and wireless feature were used to minimize possible data error induced by user discomfort due to the hardware setup's intrusiveness. For the Virtual Reality part of the experimental process, an HTC Vive Pro Eye device was used to provide the video content to each participant.

3.2. Dataset

3.2.1 Heart Rate Variability metrics

Considering essential points such as external interferences, complex calculating processes, and short time windows for data capturing, the final set of chosen Heart Rate Variability metrics to be calculated comprised SDNN, NN50, pNN50, RMSSD, MHR, and MRR.

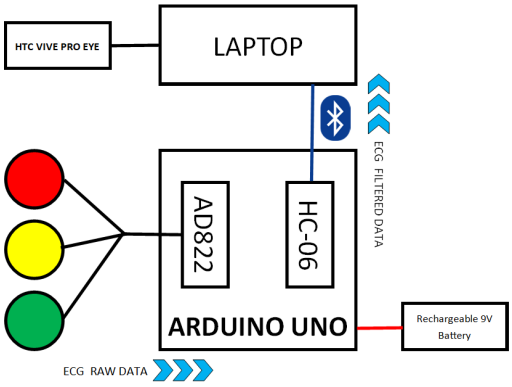

Figure 1: Diagram of the whole hardware setup.

3.2.2 Chosen Movie Clips

The used movie clip database has 64 film clips, with each arousal level being a continuous value between 1 and 7. Reducing this amount of clips by selecting those with a 3-minute duration, at maximum, turned this set into a 19-clip pool. The main objective with this selection was to end up with three clips for the Traditional Watching Experience and other three for the Virtual Reality Watching Experience, each set of 3 clips with three different levels of arousal.

3.3. Arousal Prediction Model

For each movie clip watched by each participant, the software developed in Unity and used for movie clip presentation and Heart Rate Variability metric calculation generated a single .json file with all of the raw and processed data. A python machine-learning algorithm then used these .json files to return their respective predicted arousal value. For the sake of future comparison between different experiences, three separate but structurally equal models were developed: one that received all the data, one that received only the data gathered in the Traditional Watching Experience, and another that received only the data collected in the Virtual Reality Watching Experience. A basic linear "if-then" machine learning approach would not be the best option since the relation between each data point, and the final result could be more complex. To handle this reflection, the chosen approach was to use a Neural Network, which means that the network would be organized as a set of neurons, each with inputs, outputs, and respective weight computation. The final network composition, either for the Traditional Watching Experience, Virtual Reality Watching Experience, or a combination of both,

is shown in the figure2.

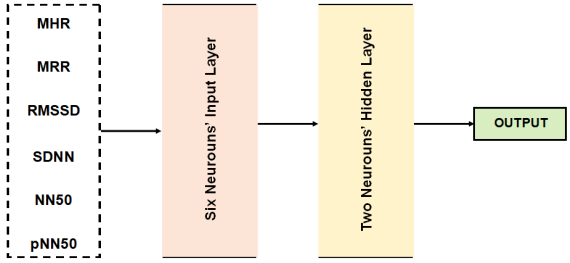

Figure 2: Final Neural Network

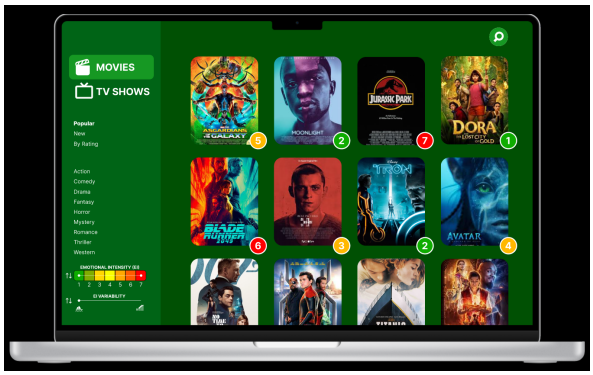

Figure 3: Streaming App Prototype

3.4. Streraming App Prototype

To understand if the proposed system would prove useful in the real world, it was decided to elaborate a medium-fidelity prototype that incorporated some features, more precisely, five, that directly used Heart Rate Variability predicted arousal to classify, filter, and order generic movies. In regards to the process of developing this prototype, it is essential to note that, even though it is understood that a complete interface designing process, with a set of previously tested prototypes, would be needed to classify this prototype as an entirely valid one, the main objective of it was not to test usability but the utility of the present features. Arguably, a shallow designing process could influence the users' reactions and the resulting analysis, but, in the sake of time and understanding that this decision was made in a late phase of the process, it was decided to keep this new data source since its purpose was considered unique and necessary. For this prototype, the design was based on a set of widely used streaming platforms, like Stremio [3] and Popcorn Time [15], while also taking some creative liberty in the new features. It is also important to point out that the chosen platform for the designing process was FIGMA due to past experiences, and the arousal was converted into Emotional Intensity, with the only intent of facilitating the understanding of the users. This prototype was then used for a user testing phase to understand if the presented arousal-based metrics would provoke any acceptance or refusal reaction next to the general audience.

3.5. Biosignal Recording

3.5.1 Participants

With a total of 20 participants, the gender distribution was seven male and 13 female individuals, and all participants were between 19 and 68 years old. All participation was voluntary and did not require any type of official commitment.

3.5.2 Procedure

Upon arrival, each participant received a full briefing about the experiment's goal and how it would be conducted. Any questions were answered, and it was ensured that the participant was comfortable with advancing with the experiment. The participant distribution was made only by attending to each participant's availability,y. Each participant filled out a form with anonymous demographic information and other vital prompts related to possible health issues, previous contact with Virtual Reality, and personal streaming habits. To eliminate the possibility of the order of experiences (Traditional and Virtual Reality watching experiences) representing a data influence, each participant watched each set of three clips in the opposite order than the previous one. Regarding the physical setup, the room lighting was lowered so that the focus of each participant could be directed to the screen of the laptop on which the experiment was run. Each participant was seated on a comfortable sofa to minimize the discomfort of being attached to sensors.

3.5.3 Results

With data being recorded from twenty different users, the final pool of results ended up with 120 .json files, one for each of the six movie clips watched by each participant. The figure 4shows the fundamental values and the predictions that resulted from the validation set (10 of the 20 participants) for the general data, Traditional Watching Experience, and Virtual Reality Watching Experience.

Regarding the consideration of all 120 data files, the model reported testing set Mean Absolute Error of 0,73, a Mean Squared Error of 0,71, and a Mean Absolute Error Percentage of 15,45%. This data resulted only from 80% of the records used for the training process (validation split of 0,2). As for the other 20%, the validation set, the final results

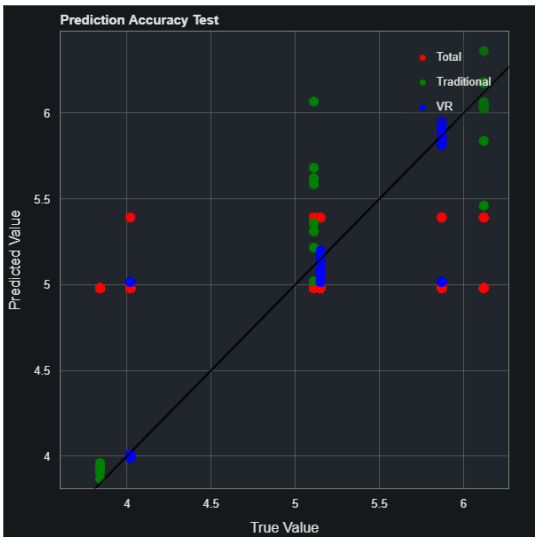

Figure 4: Arousal predictions that resulted from the validation set

were comprised of a Mean Absolute Error of 0,74, a Mean Squared Error of 0,73, and a Mean Absolute Error Percentage of 15,53%. These results show a relatively low error, but it should be taken into account that, due to the capturing window's size limitation, only three levels of arousal were targeted, which can influence the performance of this model.

When considering the Traditional Watching Experience and Virtual Reality Watching Experience models separately, each was supplied with 60 different records from the same 20 participants. Regarding the Traditional Watching Experience, the correspondent model ended up with a testing set Mean Absolute Error of 0,18, a Mean Squared Error of 0,07, and a Mean Absolute Error Percentage of 3,39%. As for the validation set, the final metrics were a Mean Absolute Error of 0,18, a Mean Squared Error of 0,07, and a Mean Absolute Error Percentage of 3,54%.

Looking into the Virtual Reality Watching Experience, the correspondent model reported a testing set Mean Absolute Error of 0,11, a Mean Squared Error of 0,07, and a Mean Absolute Error Percentage of 2,23%. Considering the validation set, the final results were a Mean Absolute Error of 0,05, a Mean Squared Error of 0, and a Mean Absolute Error Percentage of 0,90%.

One of the main objectives of this experiment was to evaluate if there is a considerable difference between the Traditional Watching Experience and the Virtual Reality Watching Experience. This could be achieved by performing a paired sample t-test with the 120 records gathered. Still, since, due to the specific capturing window's size, only arousal levels from 3,840 to 6,120 were able to be predicted, it was decided to separate this t-test into three levels: Level 1 (arousal level of 4, rounded), Level 2 (arousal level of 5, rounded) and Level 3 (arousal level of 6, round). This split was also supported by the fact that both the Traditional Watching Experience and the Virtual Reality Watching Experience were comprised of a movie clip from each of these arousal levels. For these t-tests, the considered α was 0,05 and the Df 9, since the total amount of participants for each level was 10, the same ones used for the validation of each model.

	Level 1	**Level 2**	**Level 3**
Sample size	10	10	10
Difference Mean	-0,282287	0,369446	0,321086
t	-2,00215	3,60907	4,54473
Df	9	9	9
P-value (one-tail)	0,0381438	0,00283377	0,000698215
P-value (two-tail)	0,0762876	0,00566755	0,00139643
Lower 95.0%	-0,601233	0,137878	0,161264
Upper 95.0%	0,0366586	0,601014	0,480908

Table 1: T-test result metrics for all three levels.

As can be concluded from the data in the figure1, and after taking into account that to reject the null hypothesis (the difference between the paired population means is equal to 0), the P-value (two-tail) needs to be less than α (0,05), only level 2 and 3 respect this rule and prove that there is a significant difference between the Traditional Watching Experience data and the Virtual Reality Watching Experience data. Regarding level 1, the conclusion is the opposite: there seems not to be a considerable difference between the Traditional Watching Experience and the Virtual Reality Watching Experience. This contradiction may seem confusing and counter-productive. Still, if we look into the evolution of the P-value (two-tail) from level 1 to 3, it is possible to understand that it seems to be lowering with each level. This data may lead to the conclusion that the stronger the expected arousal for a movie clip, the more accurate the Virtual Reality related predictions are. This conclusion needs, though, to be taken with a grain of salt since only three targeted levels and nine participants can provide data for the test.

3.6. User Utility Test

3.6.1 Overview

This second stage of user testing was not directly related to the first one. This means that the data acquired in the first phase was not used for this one. The main objective of this part was only to evaluate if people would find the usage of supposedly biosignal-based classification metrics in a movie streaming app ambient to be justified and worth it.

3.6.2 Participants

There was no prior participant selection process apart from availability, and the necessity of neither one being also recruited for the first phase. This guarantee was established so that each user would not be influenced by the technical part of the proposed solution since it would never be presented to its final users on that form. Eight individuals participated in this study, six female and two male. Regarding the age span, since we are talking about a streaming app and the experiment was done via a web-conference app, the younger participant was 18 years old, and the oldest was 23 years old. Procedure On arrival, each participant received a full briefing about the experiment but was not informed of the source of the biodata. They were only told that the presented streaming app introduced some new features based on heart rate-related data and that it is entirely symbolic and inaccurate. Each user was presented with the prototype to be tested. The conductor of the test followed a script previously elaborated that followed the guidelines of Google's UX Certificate to provide an external theoretical background. Each participant was asked about their will to participate, and it was preferred not to record the sessions to provide more freedom and comfort to the user. The observer and conductor of the study recurred to taking notes about each user's participation. Participants were asked for specific tasks and instructed to follow a Think Aloud perspective. At the end of the experiment, there was a direct communication window to understand users' concerns and suggestions better. Results The script for this testing phase contained a simple identifying question that prompted the user to find the presented movie with the highest Emotional Intensity score. This question led to a percentage of correct answers of 87,5%, corresponding to one incorrect answer. Regarding the wrong answer, the user overlooked one of the available movies and its corresponding Emotional Intensity score, looking instinctively at the red background and not the number itself. Every participant responded affirmatively when asked if this feature should be considered useful.

The Emotional Intensity Information Panel feature asked the user to identify any specific information they could consider less, and most importantly, 62,5% chose the number of records as the essential info point and the Emotional Intensity score as the less important one. When asked if this feature was useful, every participant answered affirmatively.

Regarding the Emotional Intensity Variation Graph, for the utility-related question, 50% of the participants answered that they did not find this feature very useful or would not use it often.

When asked to filter the presented movies by Emotional Intensity score, even tho every participant eventually succeeded and found this feature useful, some did not get it right the first time.

It was also asked the participants to order the movies shown by Emotional Intensity score, with proved to result in a unanimous conclusion that this feature, although useful, was not directly accessible to every user.

Lastly, for the Emotional Intensity Variability Filter, every user managed to fulfill the given task, but only 37,5% found this feature useful. The majority argued that it could represent too much information and a high degree of freedom that could lead to confusion.

4. Discussion

As represented in the section 1, the goal of the new approach for movie and tv show classification proposed in this document was to introduce a unique data point about this type of media directly related to the bioreaction of participants when watching it.

When trying to understand if this type of classifier would be useful for the general audience, we ought to take a look at the second testing phase, as this was the one direct opportunity to evaluate if a biosignal-based classification of movies and TV shows would, effectively, prove itself useful next to the adequate stakeholders. In a simple quantitative look, from the six different features presented to each participant, we reached an average of 72,9% when considering the one common question for all those, which aimed to understand if the participant found that specific feature useful. The only feature that did not reach the classification of "useful" was the EI Variability Filter, which did not significantly influence the general utility of biosignal-based info. At the end of each session, it was asked to each participant if they would, if asked, want to contribute with their data for this type of system regularly. Even though the majority (six out of eight) pointed out that, if ensured of their privacy and if this data were taken in a non-intrusive way, they would not have any problem sharing their measured bio reactions, the general reluctance was palpable. The final question presented to each user as if they think that they would use I regularly if their current preferred streaming app implemented a system as the one suggested. Every participant said that this new type of classification would, without a doubt, be helpful daily.

Regarding the bio-gathered data validity, it is possible to note that the model with the most considerable Mean Absolute Error Percentage is the one that merges data taken in the Traditional Watching Experience and the Virtual Reality Watching Experience, with a value of around 15%. This could lead to a general accuracy of about 85%. Still, it is essential to note that because this model

uses data gathered in different setups/situations, it must not be taken with deep concern. This model served, majorly, as a ground zero, a basis for the whole experiment, making it a lot more interesting to look into the data from the Traditional Watching Experience and the Virtual Reality Watching Experience separately. The Traditional Watching Experience and the Virtual Reality Watching Experience reported a Mean Absolute Error Percentage percentage of 3,54% and 0,90%, respectively, regarding the validation set constructed immediately before training. Due to their low values, these percentages can support that the resultant predictions are valid and can be used for movie/TV show classification. These results can also be considered real-time since the source data was taken directly from the user while watching the respective movie. They are also isolated from conscious mental processes since data was taken via biosensors and did not require any direct user data supply.

Another critical reflection is about the difference between the Traditional Watching Experience and the Virtual Reality Watching Experience and their respective data. The general conclusion is that the t-test proved this difference for the highest two of the three arousal levels to be targeted and rejected for the lowest one. This fact, combined with the consideration that the t value seems to increase considerably when upping the expected arousal value, can hint at a possible conclusion that the bigger the experienced arousal, the more precise Virtual Reality related data is. Still, this conclusion needed a more extensive set of data to be fully explored and confirmed.

When evaluating the trustworthiness of this system, the final question of the second testing phase was applied precisely to attend to this need for validation. Each participant was asked if and how, after being told that this data would instead be resultant from a set of questionnaires, this new information would influence their degree of confidence in the whole system. From a set of exciting considerations, the consensus was that the proposed approach should result in more reliable data, even though it may not be able to provide all the necessary info for a classification system by itself.

4.1. Future Work

4.1.1 Hardware Advancements

With the growth of the streaming industry, a development predicted to keep pace after the pandemic, new ways of consuming content will become gradually more frequent, and the spectrum of hardware available will become much more open and increase the audience's freedom of choice.

Virtual Reality is one of the new ways of watching movies/TV shows and one with a similar predicted path of increased popularity. With the virtualization of content and the growing appeal of Virtual Reality environments, watching movies/TV Shows is expected to become part of this new way of interacting with digital content. Another advantage, and one directly related to the approach proposed in this document, is the introduction of biosignal recording hardware in Virtual Reality equipment. These sensors are initially designed with the single purpose of being seemingly integrated with the hardware needed for a Virtual Reality experience. This makes them invisible to the users and almost reduces the voluntary participation question to a single checkbox or definition preference in a streaming app UI. One currently existing solution that could be directly applied to this situation is the HP Reverb G2 Omnicept Edition [1]. This headset already provides, apart from eye tracking, pupillometry, and face tracking, heart rate information measured directly through the provided headset. It even offers a development suite that automatically analyses Heart rate Variability and allows this info to be extracted and used by developers when developing their apps. This way, all the data needed for this solution would be directly available and represent an almost zero level of intrusion for the user.

4.1.2 Other Industries

Although the proposed system was directly aimed at the movie/TV Show streaming industry when its concept was presented, more industries could profit from using a bio-acquired Emotional Intensity score classifier. A unique look will be taken into the Marketing and Music industries to explore this possibility further.

In this case, the music industry can be directly connected to the movie/TV show industry. An Emotional Intensity analysis could be useful not only when providing the general mood of the song to the users but also to create new understandings next to the artists and producers about new releases and the intensity of the general song or even the evolution of this score throughout the music. One could argue that emotions play an even more significant role when speaking about music and its purpose, so a direct analysis of the audience and their bodily reactions could increase the proximity between the artist and the public.

Looking into the Marketing industry is one of the industries which main target is the general public and how they react to visual content. As established before, the proposed Emotional Intensity score can be a direct path to user reactions without any conscious interpretation or interference. Using this type of classification/audience reaction meter, one of the possible new approaches would be

to apply it to a user research session and directly compare the verbal feedback of each participant to their unconscious bodily reaction. This way, acquiring more data and detecting possible incoherence-related errors with verbal feedback would be possible.

5. Conclusions

This study presented a new approach for movie/TV show classification based on Heart Rate Variability. This solution started with theoretical research about possible links between emotional status, bodily reactions, and visual content provided. Once these links were justified, an Arduino-based system was developed to gather the Electrocardiogram data necessary for the Heart Rate Variability analysis. A user study with 20 participants was conducted to gather the bio-data needed to be fed to a machine learning algorithm that took twelve different Heart Rate Variability features and supplied a constant arousal value directly related to the watched movie clip. This data ended up showing a low error percentage. Lastly, a second user study was conducted, a usability test directed to a streaming app prototype that implemented some Heart Rate Variability related features, aiming to understand if a system like the proposed one could be helpful to the audience. This system proved valid, returned unconscious and real-time results, and was also considered helpful by a population of users. Future developments and expansion perspectives were also discussed.

Acknowledgements

I would like to thank my supervisor **Professor Augusto Esteves**, who handled my almost existential doubts with such finesse and was able to guide me during this long and new process. For every critical question, every direct and straight to the point answer, I can only express may deep appreciation.

To my parents I can only try to find words, because that is a hard enough task. Every bug-related scream, every existential crisis, even every meal kindly brought to me while I was spending long hours working on this project, always making sure that I was not disturbed by that attitude that meant the exact opposite than a disturbance of any kind. Twenty three years later I still do not have words, maybe I will find them in the next decades.

I would also like to thank something that is more than a single person: Theater. For almost ten years those old black planks have been my lift raft in a sea of doubts and insecurity, and it was not different in the last year. The only thing that forced my to leave my PC and take a breath were the rehearsals, and for that, I can't say enough. Art has been part of my life since I can remember and many of this dissertation's insights were made while letting art flood my brain and compartmentalize everything else.

Lastly, I want to thank Institute Superior Técnico, an institution that received my as one of its own. I never felt anything but unconditional support, well, maybe some headaches. What matters is, even when the biggest nightmares were caused by it, I always knew that a love-hate relationship is always more fun.

References

[1] Hp reverb g2 omnicept edition.

[2] SKYBOX VR Video Player on Oculus Quest.

[3] Watch videos, movies, TV series and TV channels instantly.

[4] The Truth About Test Screenings, 10 2019.

[5] Global film and video services market report 2021 - opportunities and strategies to 2030 - researchandmarkets.com, Sep 2021.

[6] Global video streaming market report 2021-2028 - researchandmarkets.com, Sep 2021.

[7] Conviva's State of Streaming Q4 2020. Technical report, 1 21.

[8] f. given i=A., given=Alexandre, f. given i=F., given=Frédéric, f. given i=X., given=Xavier, and f. given i=P., given=Pierre. Assessing the effectiveness of a large database of emotion-eliciting films: A new tool for emotion researchers. *Cognition Emotion*, 24(7):1153–1172, 11 2010.

[9] f. given i=R., given=Ramya. How to Build a Movie Recommendation System - Towards Data Science, 8 2022.

[10] L. O. Molinero. Cinema: audience emotions under the spotlight, Dec 2019.

[11] J. A. Russell. A circumplex model of affect. *J. Pers. Soc. Psychol.*, 39(6):1161–1178, Dec. 1980.

[12] M. Slater. Immersion and the illusion of presence in virtual reality. *Br. J. Psychol.*, 109(3):431–433, Aug. 2018.

[13] M. Slater and S. Wilbur. A Framework for Immersive Virtual Environments (FIVE): Speculations on the Role of Presence in Virtual Environments. *Presence: Teleoperators and Virtual Environments*, 6(6):603–616, 12 1997.

[14] V. T. Visch, E. S. Tan, and D. Molenaar. The emotional and cognitive effect of immersion in film viewing. *Cognition and Emotion*, 24(8):1439–1445, 2010.

[15] Wikipedia contributors. Popcorn Time, 8 2022.