



**PREDICTING TWITTER ACTIVITY BASED ON
PHYSIOLOGICAL DATA:
A MACHINE-LEARNING APPROACH**

By

Sara Manuela Cortinhas Pereira

Master Thesis in Modeling, Data Analysis
and Decision Support Systems

Supervised by:

Prof. Dr. João Manuel Portela da Gama

Prof. Dr. Bruno Miguel Delindro Veloso

Doutor Pedro Silva Moreira

Faculdade de Economia

Universidade do Porto

2021

Acknowledgements

I want to start by giving my sincerest thanks to those who helped me during this work. For those who had the kindness to provide me with advice, a friendly word or share their knowledge, thank you.

To Professor Doutor João Manuel Portela da Gama, supervisor of this dissertation, who guided me through this journey with his experience, admirable knowledge, patience and availability.

To Professor Doutor Bruno Miguel Delindro Veloso, co-supervisor of this dissertation, for all the support and encouragement during the periods of blockage.

To MindPorber, especially Doutor Pedro Silva Moreira, for trust me in this project and allow me to develop something meaningful. Thank you, Pedro, for all the guidance and insightful knowledge.

To Ana Martins, who shared this journey with me over the last two years. Thank you for your friendship and all the up late-nights debugging all our steps.

To all my friends for not being upset about dinner cancellations and missed conversations. Each of you gave me the strength to accomplish this.

A special thank you to Diogo, who was always on my side and believed in me in the moments I doubted the most. Thank you for all the patience, for the harshest words and for the sweetest ones. You bring so much love and light to my days.

To my family, to whom I owe most of my achievements. To my mother, for all the unconditionally support. To my sister, who is always a message away, ready to help me. To my brother Martim for giving me so much joy and not giving up asking to play a game. To my father, that is proud of me. Also, a thank you to my grandmother and Armando.

Sara Manuela Cortinhas Pereira

Abstract

Television (TV) programming continues to draw ever-larger viewers and is continually changing to adapt to new social realities. The number of people who watch a show is the most standard measure of its success. However, with the realm of social media, the number of online reactions on social media has become a more representative predictor of the show performance given the actual viewer's behaviour. Television networks are increasingly relying on social media platforms to promote their shows and engage with their viewers. This study examined whether the Twitter activity can be predicted using Electrodermal (EDA) activity, an implicit biometric measure and is connected with arousal. In collaboration with MindProber, more than 20 television programs featured on Portuguese or British television networks were analysed through the signal of electrodermal activity using data mining techniques. Six measures were extracted to meet the phasic and tonic components of the signal. The results did not indicate that using the six EDA scores is necessarily better than using the composite EDA measure. This study contributed to the growing literature on social engagement measurements through biological responses, highlighting the need to combine the latter with more show-related features. This study is a very challenging problem as (i) the desired outcome is a population-related metric while the input data is individual, and (ii) there are various external factors to the stimuli that may indirectly affect the buzz about a show.

Keywords: Television Social Engagement, Twitter, electrodermal activity, Data Mining, Ensemble Techniques

Resumo

A televisão continua a atrair um número cada vez maior de espectadores, estando em constante mudança de forma a adaptar-se às novas realidades sociais. O número de pessoas que assistem a um programa de televisão é a medida mais comum para medir o seu sucesso. No entanto, numa era dominada pelas redes sociais, o número de reações online nos meios de comunicação social tornou-se um indicador mais representativo do desempenho do programa para o atual comportamento do telespectador. As estações de televisão apostam cada vez mais nas redes sociais como forma de promover os seus espetáculos e de se envolverem com os seus telespectadores. Este estudo analisou se a atividade do Twitter pode ser prevista utilizando a atividade Eletrodérmica (EDA, do inglês Electrodermal Activity), uma medida biométrica que não pode ser controlada e que é associada a excitação face a estímulos. Numa colaboração com a MindProber, mais de 20 programas de televisão apresentados em redes de televisão portuguesas ou britânicas foram analisados através do sinal da atividade eletrotérmica, utilizando técnicas de mineração de dados (também conhecida pelo termo inglês data mining). Foram extraídas seis medidas para satisfazer as componentes fásica e tónica do sinal. Os resultados não indicaram que a utilização destas últimas é necessariamente melhor do que a utilização do sinal Eletrodérmico composto. Este estudo contribuiu para a crescente literatura sobre medições de envolvimento social através de respostas biológicas, salientando a necessidade de combinar as últimas com características mais relacionadas com o programa de televisão. Este é um problema desafiante, pois: (i) o resultado desejado é uma métrica relacionada com a população enquanto os dados de entrada são individuais e (ii) existem vários fatores externos aos estímulos que podem afetar indiretamente o número de conversas que o programa vai gerar nas redes sociais.

Palavras-chave: Televisão, Redes Sociais, Twitter, atividade Eletrodérmica, Mineração de Dados, Modelos Múltiplos

Contents

Acknowledgements	iii
Abstract	iv
Resumo	v
1 Introduction	1
1.1 Motivation and Context	2
1.2 Problem formulation and objectives	3
1.3 Dissertation Structure	4
2 Related Work	5
2.1 Television Success Measures	5
2.2 Electrodermal Activity (EDA)	8
2.3 Twitter	12
2.4 Predicting Television Show Success	14
3 Case Study	18
3.1 Business Understanding	18
3.1.1 Procedures & Participants	18
3.1.2 Monitoring Sessions (Stimuli)	19
3.1.3 Bussiness Core	19
3.2 Data Analysis	19
3.2.1 Episode Level Analysis	19
3.2.2 Electrodermal Signals	21
3.2.3 Twitter	26
3.2.4 Data Input	28
3.2.5 Software	30
3.3 Modelling	30
3.3.1 Modeling Techniques	31
3.3.2 Model Training using EDA Components	31
3.3.3 Model Training using the composite EDA signal	38
3.3.4 Models Assessment	39
3.4 Evaluation	40
3.4.1 Performance Metrics	41
3.4.2 Results and Conclusions	42

4	Conclusions	45
4.1	Limitations	47
	Bibliography	48
	Appendix	i
A	Distribution of Twitter Activity per episode	ii
B	Inter-subject Correlations	iv
C	Mean and Standard Deviation between participants for airtime	x
D	5-Folder Cross Validation Results for XGBoost Model using EDA Components	xvi
E	5-Folder Cross Validation Results for XGBoost and MLP using the composite EDA Signal	xix
F	Visualisation of the errors for MLP model	xxii
G	Visualisation of the errors for XGBoost model	xxv

List of Figures

2.1	Decomposition of EDA data into tonic and phasic components	10
2.2	Properties of SCR	12
2.3	Tree Ensemble Model	16
2.4	Structure of MLP Neural Network	17
3.1	Electrodermal Activity across episode duration (Episode E14)	22
3.2	Decomposition of the EDA Signal	23
3.3	Matrix of Spearmans's correlation between participants for the raw signal of electrodermal activity	24
3.4	Twitter volume across episode duration (Episode E14)	26
3.5	Distribution of Minute-to-Minute Differences in Twitter Activity (divide in two groups according to the activity volume)	27
3.6	Minute-to-Minute Changes in Twitter Activity and EDA Signal in z-scores (Episode E14)	29
3.7	Ensemble Architecture	31
3.8	Implementation of 5-Folder Cross Validation	34
3.9	Feature Importance for XGBoost	41
3.10	Prediction of Twitter Activity	43
4.1	Distribution of Twitter Activity grouped by volume of tweets	ii
4.2	Matrix of Spearmans's correlation between participants per episode	iv
4.3	Mean and Standard Deviation between participants for airtime	x
4.4	Prediction of Twitter Activity for MLP Model	xxii
4.5	Prediction of Twitter Activity for XGBoost Model	xxv

List of Tables

2.1	Description of Electrodermal Activity Measures	11
3.1	Television Show information	20
3.2	Mean inter-subject correlations for the raw signal of electrodermal activity for the whole episode	25
3.3	Properties of the studied data sets	29
3.4	Results of the 5-fold cross-validation run: Average and standard deviation of the MAE for XGBoost Model using EDA Components	36
3.5	Results of the 5-fold cross-validation run: Average and standard deviation of the MAE for MLP Model using EDA Components	36
3.6	Default Values, grid search interval and the chosen values for XGBoost Model .	38
3.7	Default Values, grid search interval and the chosen values for MLP Model . . .	38
3.8	Results of the 5-fold cross-validation run: Average and standard deviation of the MAE for XGBoost Model using EDA Signal	39
3.9	Results of the 5-fold cross-validation run: Average and standard deviation of the MAE for MLP Model using EDA Signal	39
3.10	Prediction Errors for XGBoost and MLP	42
4.1	Results of the 5-fold cross-validation run: Average and standard deviation for XGBoost model using C2	xvi
4.2	Results of the 5-fold cross-validation run: Average and standard deviation for MLP model using C2	xvi
4.3	Results of the 5-fold cross-validation run: Average and standard deviation for XGBoost model using C3	xvii
4.4	Results of the 5-fold cross-validation run: Average and standard deviation for MLP model using C3	xvii
4.5	Results of the 5-fold cross-validation run: Average and standard deviation for XGBoost model using C4	xviii
4.6	Results of the 5-fold cross-validation run: Average and standard deviation for MLP model using C4	xviii
4.7	Results of the 5-fold cross-validation run: Average and standard deviation for XGBoost model using R2	xix
4.8	Results of the 5-fold cross-validation run: Average and standard deviation for MLP model using R2	xix

4.9	Results of the 5-fold cross-validation run: Average and standard deviation for XGBoost model using R3	xx
4.10	Results of the 5-fold cross-validation run: Average and standard deviation for MLP model using R3	xx
4.11	Results of the 5-fold cross-validation run: Average and standard deviation for XGBoost model using R4	xxi
4.12	Results of the 5-fold cross-validation run: Average and standard deviation for MLP model using R4	xxi

Chapter 1

Introduction

In the early 20th century, Television (TV) was designed and conceived to be a medium for entertainment, education, or informative communication, and quickly became a popular broadcast medium. Nowadays, it is hardly a day that goes by without people watching television programs or video content. Television is well established as a daily routine in consumers' life. The characteristics of TV permitted the emergence of an entire media landscape and new business and consumption models to emerge. This transformation process started with a change in the way media is delivered - an increased number of channels is nowadays available spanning broadcast networks, cable, and online streaming companies that produce television contents of varying quality and genres. Moreover, the introduction of new technologies and the realm of social media revolutionized the way people experience and perceive television. People now can voice their thoughts and opinions on social media platforms, and broadcasters are eager to absorb it and deliver the content of higher quality.

The prevalent measuring methods for television viewing rates are mostly given by the number of viewers that watched an episode at any given moment or overall. Besides that, given the aforementioned changes in media consumption, the social online engagement of the viewer with the TV shows has become another important success measure. However, while contemplating the new dynamics and television experiences, it is important to keep in mind that certain conditions of television remain the same. It is still important to know what makes television shows good and the drivers of that success to be able to understand the reasons for why a person watches television. The strength of the creative content, or emotional payoff, is undoubtedly seen as the main factor for a program to resonate well with viewers. Hence getting the most out of the creative content to maximize TV viewership and social engagement is very critical for broadcasters and show producers.

This ongoing study focuses on determining whether social engagement, as expressed by Twitter activity, can be predicted by real-time viewer responses measuring through electrodermal activity which are less corrupted by prejudices in reporting. The changes in sweat gland activity can be analyzed as an indicator of emotional arousal given the link to different cognitive processes. Ultimately, the aim is to identify moments of the content that contribute to a low level of engagement that is followed by the person shift to another program or stop watching television, and conversely, the moments of a higher level of engagement that attracts new viewers, during

broadcast time.

To accomplish this study, machine learning techniques will be applied by utilizing the audiences' free writings on Twitter and the electrodermal responses, comprised of time series data. The study of predicting television success is mainly dominated by linear models, that have been able to perform relatively well in this task, however, due to the increase in complexity of data, advanced methods, such as Data Mining and Machine Learning, are becoming relevant in this field. We aim to contribute by examining and comparing the performance of Artificial Networks models and Ensemble Models in providing an objective view of how an audience will respond to a presentation.

The received dataset refers to thirty-one different episodes across different television shows and it was provided by a real-time, physiological and behavioural analytics media testing platform, MindProber.

1.1 Motivation and Context

Until recently, television viewing was associated with a passive experience to entertain, inform, or educating a person or group of people through television in the living room. However, the appearance of mobile devices and the realm of social media brought new dynamics to media consumption and changed the way people experience television. Besides that, people can watch television regardless of where and when through their mobile advice, and they are also interacting with the show by commenting and sharing their thoughts about it on social media or interactive apps.

Out of this transformation and expansion, an increasing number of television shows of different genres and varying quality can be found on traditional networks, cables, and online streaming companies, such as Netflix, HBO, and Amazon. This means new opportunities and challenges for networks, producers, and advertisers, who seek to know what makes television shows good and which ones will lead to more engagement from the viewers.

This study is part of a collaboration with MindProber, a startup that offers innovative and modern tools for media content, powered with biometrics in real-time for large scales environments. The small and non-intrusive biometrics sensors allow for the collection of human body reactions from hundreds of viewers that are used to measure the viewers' engagement with media content and predict audience churn. Additionally, their mobile app allows for the recording of declarative answers about the content.

This project aims to predict Twitter activity based on time-series of behavior and physiological data, namely electrodermal activity. Audience churn is an important indicator of the popularity of shows and how people react to media content. Identifying the most engaging moments during the broadcast that increase the number of viewers or, conversely, identifying the moments that drive people to stop watching the content is undoubtedly valuable, and it influences several strategical decisions of TV makers.

1.2 Problem formulation and objectives

The success of a television show is either measured by the number of viewers watching it or through viewers' social online engagement. The main goal is to be able to predict the changes in Twitter volume across program sections using electrodermal (EDA) activity measures - a measure sensitive to emotional arousal.

All data was collected and provided by MindProber. The available data comprises two independent variables. The biometric timepoints for electrodermal activity - obtained by monitoring the panelists of spectators while they are watching TV programs and while wearing the psychophysiological sensors. The outcome is a time series per participant reflecting the level of sweat gland activity.

In addition, Twitter activity is also provided in a time series displaying all the posted tweets during the broadcast time of the show. The selection of the tweets was done by extracting all tweets that contained specific hashtags, usually the official hashtag created by television networks. The volume of Twitter activity may be an indicator of how people are engaging with the show, reflecting their thoughts and opinions for specific events.

Briefly, the data comprises the biologically based responses of hundreds of participants to the content during the visualization of specific television shows, for example reality shows and sports events. The total number of tweets represents the reality that we seek to understand and predict.

The current study is aimed at determining whether EDA measures can prove to be of predictive power behind other measures of Television show quality and explain more variance in Twitter activity than other behavior variables, or demographic features. And second, we aim to determine whether population-level success, here expressed as change in Twitter volume activity during airtime, of a selected range of Television shows, involving 31 episodes from several shows, could be predicted using electrodermal activity as expressed in changes in sweat gland activity.

More precisely, at the end of this project, we intend to answer the following research questions that are largely related and concerned with the elements that most contribute to the outcome value we want to predict. We are primly interested in knowing to what extent it would be possible, based on EDA measurements, to identify moments of low and high arousal, here indicative of social engagement. Can sweat gland activity be used to predict Twitter activity? Moreover, we want to examine whether the behavioural variables taken in isolation from the EDA signal can be used to predict changes in Twitter volume across each program's minute, and to what extent they are better predictors than the individual signal. The contribution and influence of each component is also a point of interest that will be further analyzed.

In various statistical fields, the sample size is an important factor and it may have direct impact on the results. Since for each episode we have data for several participants, we aim to understand the impact on the prediction model of changing the sample size in terms of participants.

The changes in Twitter activity give a population-level behavior whereas the electrodermal activity is tied to each individual. Considering that these two variables are in different populations levels, we hypothesise whether we can aggregate the individual points to obtain a single measure that would be representative of the population arousal. Thus, we aim to access the contribution of aggregating the biometric points over the signals of each individual as input data.

1.3 Dissertation Structure

The remainder of this project is organized as follows: Chapter 2 begins by outlining the relevant research in the area with a brief review of key points to construct a relevant literature review. This theoretical framework includes an initial contextualization and exploration of the television success measurements and the drivers of such success. The contributions of electrodermal activity researches with a focus on viewers' engagement were introduced. Relevant findings in the alliance of social media with television content are also included, namely the power of Twitter in gain viewers' engagement. Chapter 3 is initialized with a contextualization from a business perspective, highlighting the procedures to collect the data to be used and the desired outputs. In a second stage, the dataset in use is described and explained through the different variables it comprises, following by a data exploratory analysis. The data preprocessing and the assumptions taken to prepare the data for the next phase are detailed. The software chosen is also presented. The chapter continues with the modeling phase where are provided the models' descriptions and parameter settings. Next, the results are presented and analyzed. Finally, conclusions and reflections are summarised in Chapter 5.

Chapter 2

Related Work

This chapter was designed to explain the conceptual framework by addressing the main concepts to consider during the entire work. This theoretical framework is divided into four key points. First, we open with the main television features by assessing the historical measurements of TV success and their drivers. Second, we then introduce the chosen neurophysiological method. Third, we further include the study of social media activity, particularly Twitter, as a predictor of engagement. Finally, we end by discussing the prediction of TV success determined by the chosen measures.

2.1 Television Success Measures

Today, Television is well established in our homes by providing numerous television (TV) shows of varying nature. Commercials, or advertisements, are a fundamental element when it comes to income. The more the advertisers, the higher the income. On the other hand, traditional Television faces the challenging competition of online streaming companies, such as Netflix, Amazon, and HBO. In this context, it becomes undoubtedly essential to produce quality programs and reach the maximum audience.

The first attempt to measure TV audiences dated 1950, and back then, the data available was of poor quality. The first models used information collected by phone surveys (Hill, 2014), where viewers were asked what channel they were turned in and personal information. This technique was then refined by introducing metering devices that allowed recording what was on the TV and information about who was watching it. Later on in the 2000s, it was further improved with the appearance of devices such as set-up boxes, video-on-demand (VOD), cell phones, and social media that result in a new type of data – more detailed data. New measurement techniques emerge as a product of this digital data. The challenge now is to evaluate the impact of TV on viewers' attention and their impressions while watching a TV show.

The growth of new technologies has brought a profound change in the way media is consumed. The wide adoption of digital mobile devices, such as tablets, smartphones and laptops, allows people to watch Television regardless of where or when. However, this digital era has not only changed the way people watch Television but also the way they experience it. With the rise of social media and interactive apps, the second screen phenomenon, defined as people

commenting and sharing their thoughts about the show with others while watching the shows, is emerging in a meaningful way. This phenomenon is part of a broader concept of Social Television (see, for example, Mantzari, Lekakos, and Vrechopoulos (2008)) and encourages a deeper level of interaction between viewers.

These changes in the ways that viewers use Television and the ways such use is evaluated resulted in new measures of audiences, such as exposure and engagement, that holds value (Napoli & Kosterich, 2015). The traditional measures of audience exposure become increasingly unfit as big data provided by social media technologies grants new ways that represent more deeply the audience behaviour. Specifically, valuing passion viewers over only on the number of “eyeballs” (Arvidsson & Bonini, 2015). This implies that when evaluating a television show, besides having many viewers, it is also important to have passionate and engaged viewers capable of influencing others. Indeed, both acknowledge social television analytics as a secondary, supplementary measurement of audiences. It is the integration of the two in measuring audience that proves to be most valuable, not replacing the traditional exposure-focused audience ratings with the new ones.

Much industry-driven research investigates the link between social media and TV programming, with particular incidence on audience behaviours of viewing and interacting with more than one screen and further reinforcing the idea of social television analytics as a complementary measurement. Nielsen granted the first research connecting Twitter activity with prime-time ratings (Nielsen, 2013). Specifically, this study undercovers a two-way influence relationship as it shows that Live TV ratings trigger more Twitter engagement. Conversely, increased Twitter activity can produce significant changes in Live TV ratings. Later research from Nielsen showed that a 10 % increase in Nielsen Twitter TV Ratings (NTTR) results in a 1.8 % increase in the +7 viewership metric (Nielsen, 2014).

Pynta, Seixas, Nield, and Hier (2014) in their study about the effects of second-screen iteration in real time showed that second screen activity positively impacts viewer engagement in a cumulative and sustained way over the longer term.

This has made social television data fascinating and worthy for Television when measuring audience value (Kosterich, 2016), as social audience activity can lead to a different direction about the type of content audiences most consume, emphasising influence, engagement or clout as drivers of valued audiences. In fact, as demonstrated in Kosterich (2016), this shift can provide a new and significantly different perception of which type of genres might become more popular or prominent and what causes it. Their findings also detach the potential of social television measurements to promote content diversity in television production.

Taken together, the above investigations suggest that the retention or cancellation of a television program may depend on factors that go beyond traditional audience ratings and including the presence of the show on social media dictated by the number of tweets, likes or shares (see, for example, Hayes (2015)).

Considering those as mentioned above, traditional ratings focus on exposure metrics, such as ratings and shares. In contrast, social television analytics recognises audience engagement with television shows, measured through social media activity, as a source of audience value. Other aspects of audience behaviour are capture by these new metrics.

The study of TV audiences has always been considered complex (Patelis, Metaxiotis, Nikolopou-

los, & Assimakopoulos, 2003). Many studies on predicting TV audiences have used more conventional features, such as genre, program content and airtime. Other approaches also include information related to the cast and their popularity to predict audience ratings (Fukushima, Yamasaki, & Aizawa, 2016). However, it is the effectiveness of the creative content that is recognised as the main driver of Television shows success (Doughty, Rowland, & Lawson, 2012; Schirra, Sun, & Bentley, 2014; Wohn & Na, 2011). The shows must be able to emotionally capture the attention and interest of the spectator (Gregersen, Langkjær, Heiselberg, & Wieland, 2017). When exposed to audio-visual presentations, our senses are stimulated, and it results in biologically based responses (e.g., skin conductance, heart rate, blood pressure). Thus, it is of particular interest to understand the drivers of such emotional responses and preferences and how can we access and measured them.

The traditional approach of measuring the spectators' responses to the content is known as interrogation or non-biologically (e.g., self-report questionnaires, moment-to-moment ratings, focus groups, surveys) and usually involves introspective reflections about the emotions experienced. The use of these methods become popular as they are cheap, quick, and relative simple to analyse; also, they provide some intuitions about viewers attitudes and preferences (Venkatraman et al., 2015).

However, such procedures are associated with high bias, lack of granularity and other vulnerabilities. Firstly, they suffer from the inability to capture the biological responses to the content. Furthermore, as these measures are post-hoc self-examinations from a prior stimulus, they can be contaminated; indeed, much of peoples' emotions and behaviours invoke processes that they are not aware of, which will result in subjective and incomplete opinions (Christopoulos, Uy, & Yap, 2016; Peacock, Purvis, & Hazlett, 2011; Shestyuk, Kasinathan, Karapoondinott, Knight, & Gurumoorthy, 2019; Venkatraman et al., 2015). Furthermore, the respondent may be resistant to share his/her honest emotions and feelings due to social desirability concerns, especially for delicate matters (Poels & Dewitte, 2006). Moreover, people may feel the obligation to respond in a certain way. Finally, on the implementation side, they are usually associated with intrusive, unnatural and distracting environments.

In more recent years, there has been a focus on biologically based responses techniques, also called autonomic information, to evaluate the quality of the content. The bodily reactions that follow emotions cannot be controlled, so measuring autonomic responses can overcome the limitations of self-report techniques. Moreover, these alternatives are more resistant to bias and allow to assess viewers' responses on a moment-to-moment basis accurately. In addition, they can capture the viewers' unconscious perceptions or emotions (Christopoulos et al., 2016; Poels & Dewitte, 2006). Nevertheless, such signals can be affected by external sources and so require close monitoring. Such techniques include eye-tracking procedures, electroencephalography (EEG) measures and electrodermal (EDA) activity.

The most common metric to measure the responses to social-media interactions is engagement, frequently used in the media. Although there could be others ways of defining it, in the cognitive neuroscience field, it is usually defined as "when a stimulus elicits a sense of personal relevance" (Pynta et al., 2014). The present study employs this metric in the subsequent analysis.

Most of the studies that depend on biologically based responses predict the success of a TV show through two steps: first by studying the viewers' engagement and then using it to predict

TV audiences. In order to create the illusion of a natural environment, some investigations have used video tools (*e.g.*, RGB video camera) to assess the engagement level (Hernandez et al., 2013). However, given the ambiguity of some behaviours, an extra effort is needed to make accurate predictions. Other biosensors such as EEG measures – related to viewers’ brain activity (see, for example, Vecchiato et al. (2010); Venkatraman et al. (2015)) and EDA activity (see, for example, Gregersen et al. (2017)) have been successfully applied to the prediction of viewership. Indeed, the study of brain activity in response to creative content using such measures has been a matter worth curiosity.

2.2 Electrodermal Activity (EDA)

Many researchers have attempted to detect the brain activity signals associated with an increase in attention, memory, or emotional engagement during the observation of audio-video contents. The nervous system, of which the human brain is part, can be anatomically divided between the central and the peripheral nervous central. The latter comprises the somatic, enteric and autonomic nervous systems. Central measures (electrical, magnetic, metabolic, hemodynamic) have contributed, in recent years, to more accurate brain activity functional maps. In turn, indicators of mind expressions are reached through somatic and autonomic measures (Sequeira, Hot, Silvert, & Delplanque, 2009). Electromyography and eye-tracking are examples of somatic peripheral electrical measures, and autonomic pupillometry, electrodermal activity or cardiovascular are classified as autonomic peripheral electrical measures.

Since the 1980s, Autonomic electrical recordings have been interpreted as a signal of emotion, attention, decision-making or unconscious detection (Sequeira et al., 2009) and further studies have sustained the role of the Autonomic nervous system (ANS) as the main way of stimulation of emotions. Here, we are particularly interested in the sympathetic and parasympathetic systems that are part of the ANS. In most cases, by balancing their activity, there is a regulation of the states of arousal that occur in the body. The sympathetic nervous system termed the “fight-or-flight system” supports the motor action and the body’s automatic responses to stimuli. It is part of emotional, cognition and behavioural experiences. An increase in sympathetic activity is manifested by involuntary increases in heart rate, blood pressure, eye blinking and sweating, to name a few physiological responses. The described physiological responses to an external stimulus are known as Biometrics.

Even though hemodynamic measurements of brain activity allow for accurate and precise identification of activated brain structures during the stimulus, the lack of time resolution due to human biological functioning makes Magnetic Resonance Imaging (MRI) unfit to provide sub-seconds activity of the brain dynamics. The popular EEG techniques overcome this limitation as it provides temporal resolution in the millisecond range. However, it falls short in assessing the deep structures responsible for the generation of emotional processing in humans by using superficial EEG electrodes (Vecchiato et al., 2010). Therefore, reconstructing the viewer’s emotional engagement from indirect signs is of utmost relevance. Indirect signs of the emotional process could be changes in sweat activity or heart rate variability. These physiological signals are indicators of the viewers’ emotional state and enable us to evaluate the level of the autonomic

nervous system. For the present work, electrodermal activity, as a measurement of arousal, will be used to evaluate the emotional responses of the audience and their engagement with the TV program.

The motivation behind the use of EDA technology come from different considerations. First, the equipment used is relatively cheap compared to more sophisticated imaging methods and provides an unobtrusive measure of audience activation. Moreover, its scalability makes it possible to obtain data from several respondents in a short period with minimal supervision. The ease of use and being 100 % implicit and discreet also make EDA a reliable marker of emotional arousal. Additionally, the subjects have no power over their EDA, which allows us to overcome the limitations of self-report. Another advantage that is associated with electrical autonomic measures, where EDA measures are contained, is that it is possible to connect the changes in any of these measures with functional consequences (Sequeira et al., 2009).

However, such measures are only indirect indicators of brain activity, specific indicators of activation or arousal, making their main disadvantage. In fact, arousal is recognised as an essential dimension of emotional responses, alongside valence (Posner, Russell, & Peterson, 2005). Entertainment is not only associated with positive valences - such as joy and positive experiences through content, but also with negative valences - such as empathetic sadness, that can lead to pleasure. Moreover, although arousal does not represent the specific emotions during viewing either the emotional valence, it indicates how the content was experienced (Gregersen et al., 2017; Venkatraman et al., 2015).

According to Leiner, Fahr, and Früh (2012), the use of EDA as a measure of the arousal of the audience of media content, or as with any psychophysiological measure, may be limited when considering the following. First, viewers get used to media content, and as a result, the responses get weaker as they experienced similar stimuli. Second, another factor that may produce different reactions among individuals is the familiarity with the stimulus, as it can lead to associations from past experiences. Not to mention that there may be other personal characteristics that shape the responses to a stimulus.

One of the factors that trigger engagement is the emotional payoff that the viewers feel while watching television (Peacock et al., 2011) and so successful television programs are the ones that emotionally connect with the audience. In addition, several studies highlighted the importance of emotions in other cognitive and behavioural responses in advertising processes (Poels & Dewitte, 2006).

According to Gregersen et al. (2017), there are three distinct levels associated with emotional engagement in TV. The first level concerns the primary responses, the immediate emotions experienced (for example, happiness or melancholy). The way they react to such impulses leads to the second level, the “short-term re-appraisal by way of meta-cognitive and meta-emotional processes, which regulate emotion”. At a later stage, it is evoked a third level implicating a re-appraisal based on social media reactions. Some studies associated the increases in arousal during broadcasting to the primary responses (Christopoulos et al., 2016; Gregersen et al., 2017).

The skin is the largest organ of the human body and the central canal that connects the organism to the environment. It plays a key role in various processes but especially in emotional communication, which means that the skin is easily excited. In particular, changes in sweat gland activity reflect the intensity of our emotional state, representing particular events with a specific

significance for individuals, usually in the emotional and attentional field (Sequeira et al., 2009). These changes in skin conductance were termed electrodermal activity.

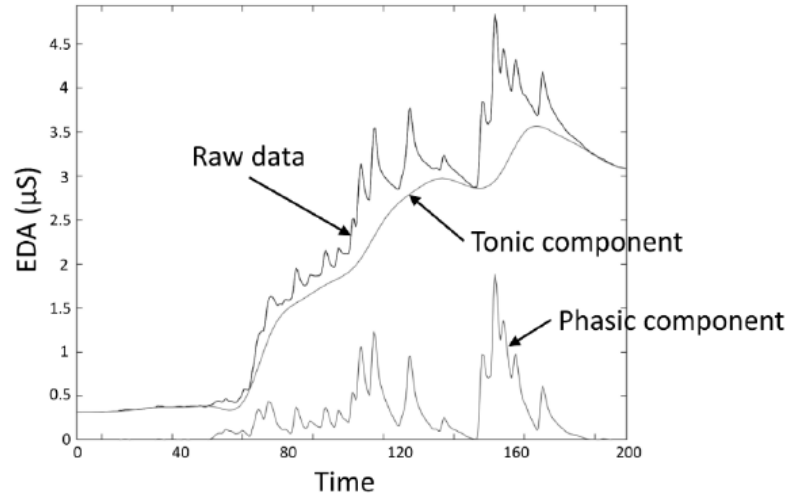


Figure 2.1: Decomposition of EDA data into tonic and phasic components (taken from Posada-Quintero and Chon (2020))

EDA is essentially controlled by the autonomic nervous system, or more precisely, by the sympathetic nervous system, which influences the production of sweat glands (Leiner et al., 2012). Concretely, the higher the activation of the autonomic nervous system, the more sweat secretion it will be produced and, as a result, a high level of skin conductance; given that the increase in activation of the autonomic nervous systems is an index of arousal, skin conductance proves to be feasible as a measure of arousal (Poels & Dewitte, 2006). Hence, skin conductance measurements allow us to indirectly measure the arousal of the sympathetic autonomic nervous system associated with emotion, cognition and attention (Critchley, 2002; Gregersen et al., 2017; Venkatraman et al., 2015). Moreover, Moreira, Chaves, Dias, Costa, and Almeida (2018) aligns with this research showing that variations of measures from the electrodermal system are directly associated with self-report arousal.

The relation between the generation and level of electrodermal activity with specific brain areas closely related to emotional and motivational behaviours is well established. These specific areas are the hypothalamus, limbic system (amygdala and hippocampus, cingulate gyrus) and frontal cortical areas (Christopoulos et al., 2016; Sequeira et al., 2009) with a particular focus for the hippocampus (with a primary role in learning and memory) and amygdala (deeply involved with emotions and other reactions to stimuli). Specifically, as stated in Richardson et al. (2020), stimulation of the amygdala means the activation of the adrenal medulla, which produces the adrenaline hormone and increases the function of the autonomic system activity, which means generation of skin conductance responses. In a similar way but with lower amplitude, stimulation of the hippocampus could also generate sympathetic responses in the form of sweat secretion.

The ongoing studies support this idea of a correlation between EDA and central psychological phenomena. Thus electrodermal activity can be used as an indicator of arousal, fractions of

Group	Measure	Definition
Tonic measures	Skin conductance level (SCL)	Tonic level of electrical conductivity of skin
Phasic measures	Change in SCL	Gradual changes in SCL measured at two or more points in time
	Frequency of NS-SCRs	Number of SCRs in absence of identifiable eliciting stimulus
	SCR amplitude	Phasic increase in conductance shortly following stimulus onset
	SCR latency	Temporal interval between stimulus onset and SCR initiation
	SCR rise time	Temporal interval between SCR initiation and SCR peak
	SCR half recovery time	Temporal interval between SCR peak and point of 50 % recovery of SCR amplitude
	SCR habituation (trials to habituation)	Number of stimulus presentations before two or three trials with no response
	SCR habituation (slope)	Rate of change of ER-SCR amplitude

Table 2.1: Description of Electrodermal Activity Measures. Retrived from Moreira et al. (2018)

attention, information processing and emotions (Critchley, 2002; Leiner et al., 2012; Vecchiato et al., 2010). In a nutshell, stimuli or tasks that demand viewers’ attention usually increase EDA responses as a result of the link between arousal and attention, that share neural properties – using EDA as an indicator of attention is reliable considering that attention is activated through stimuli that induce arousal (Critchley, 2002), as well as stimuli that involve emotional meaning are known to result in EDA signals.

The signal of EDA comprises two phases: the tonic and phasic components - see Figure 2.1 for visual comprehension. The tonic phase, also known as Skin Conductance Level (SCL) or Electrodermal level, is related to the slow changes of sweat gland activity. This component is a way to point out the general level of arousal. On the contrary, the phasic component can be seen as responses to the stimuli, commonly known as skin conductance responses (SCRs) or Electrodermal responses. The latter is then of most relevance as it identifies particular events (Posada-Quintero & Chon, 2020) that lead to arousal and so to the viewer’s engagement.

Besides the said components, electrodermal responses can be described through their amplitude, their latency, their duration and their surface (Leiner et al., 2012; Sequeira et al., 2009). In particular, electrodermal amplitude (*i.e.*, the time difference between the stimulus and response) and response latency are direct indexes of arousal when watching a video-audio content (Venkatraman et al., 2015). As arousal increases, disregarding valence, the amplitude of skin conductance responses increases linearly (Sequeira et al., 2009). Sweat glands, specifically eccrine glands, are

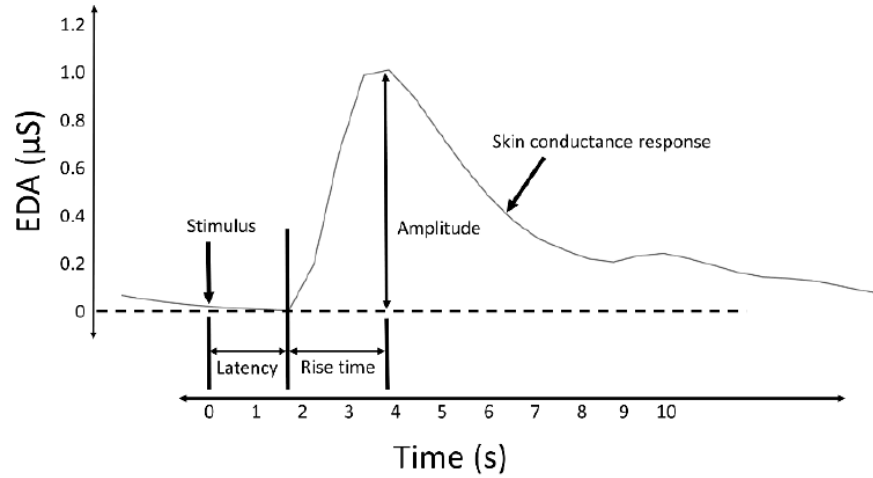


Figure 2.2: Properties of SCR (taken from Posada-Quintero and Chon (2020))

the ones responsible for the generation of SCR and are primarily found in the palmar and plantar hypodermis (Christopoulos et al., 2016; Poels & Dewitte, 2006; Sequeira et al., 2009). Although the main job of sweat glands is thermoregulation, because they are sympathetically innervated, they can also be associated with emotional responses, frequently designed as emotional sweating.

A description of the constituting measures of Electrodermal activity is presented in Table 2.1. In addition, an example of the components of the Skin Conductance Responses when a specific stimulus occurs is also given in Figure 2.2.

2.3 Twitter

Twitter and Facebook are widely used Social Media by viewers to voice their opinion about the television program they are watching. However, it is on Twitter that the most substantial and real-time discussions about television shows are happening.

Twitter is a micro-blogging platform that allows users to publish short messages. Twitter users can direct their messages for a person of interest by simply adding “@” in front of that person’s username and search for specific topics. Twitter itself recognised a ‘special connection’ between television and Twitter (Inc., 2011). Their analysis highlighted the best Twitter streams for marketing purposes as television shows displaying official hashtags on-screen encourage ‘trending topics’ or read live tweets on the air. ‘Trending Topics’ is a Twitter service that shows words, phrase or hashtags that is becoming popular very quickly as users use it in their posts. It is common for television programs to be a Trending Topic, and when occurring at airtime, it can be read as an indication that people are tweeting while watching the show.

The use of Twitter during live Television to share their opinions and express their feelings with others watching the same show has become increasingly popular among shows’ fans. This act - known as “live-tweeting” - caught the attention of television producers. Twitter is seen as a complementary tool for broadcasters that may be used to promote engagement between the audience by driving conversations about their programming and attracting additional viewers to

tune in to programming (Nielsen, 2013). As aforementioned, Twitter activity became an accepted and innovative measure of the success and impact of television shows.

In fact, Wakamiya, Lee, and Sumiya (2011) describes a new type of integration between the television makers and general audiences resulted from this new pattern of audience behaviour. That is, television makers are interested in audiences' opinions and thoughts about their shows' content. For example, networks created an official account for the shows to interact with fans through online polls. They also display an official hashtag to promote live tweeting. Hashtags are words or phrases that are preceded by a hash signal (#). Conversely, people are eager to participate in their favourite shows by expressing and sharing their opinions and feelings with other people who also watch the show or directly to the production and cast of the show. This two-way channel seems to be prompting engagement between both sides. Owing to social media, it is easier to gather broader audiences rating with no additional recruitment's costs, overcoming the limitation of the traditional television ratings.

Much research was conducted in this realm of social media to understand the motivations behind live-tweeting and provide extended means to analyse viewers' sentiment and engagement. For example, in an attempt to analyse the content of tweets for televised political and entertainment programs, Wohn and Na (2011) categorised Twitter messages according to four different types - attention, emotion, information, and opinion, and showed that 60 % of live tweets falls either in the emotional or opinionated types. Moreover, their findings uncovered a strong correlation between the type of messages posted and the program's content, suggesting that people post their reactions to what they are seeing.

Further, Doughty et al. (2012) also pointed out the television broadcast content as the contextual driver for second-screen activity. Viewers usually comment on or discussed topics related to the actors and the content of the narratives in their social network platforms. This research that tries to understand the nature of Television shows' audience that interacting with Twitter as a second screen activity also anticipates leverage in such studies. The information and the conversations on social media such as Twitter can potentially expose audience characteristics and engagement features that further provide valuable information for the television makers. Their findings showed that the proportion of networked audiences involved in discussions with a friend is a resultant of the content and tone of the television programs.

Auverset and Billings (2016) also investigated the motivations for second screen use on Twitter for the popular television show *The Walking Dead*; the results showed that social television activity is a hybrid experience involving cognitive efforts not only during the airing of a show but also before and after it. For example, tweets related to upcoming events could represent the potential enthusiasm and expectations of the viewers. Conversely, viewers try to develop deeper conversations about what they just saw after the show's end.

Schirra et al. (2014) focused on the quantitative study of live-tweeting behaviour related to an entire season of a television show. This study indicates that this act is a complex social process, depending on many personal considerations. Plot twists or story arcs that cause surprise or shock were found to be live-tweeting triggers. Viewers typically turn to Twitter when they are emotionally connected, sharing their sadness or euphoria for specific events. In addition, the authors found, behind the personal benefits from posting on Twitter, the sense of connectedness with a broader audience as the main driver to use Twitter. People tend to post more when

watching Television alone or have no friends with the same tastes. The live-tweeting community allows for this sharing and access to other people’s reactions to critical moments.

Indeed, this study shows that posting tweets may not be the first instinct when live-tweeting, but instead, people are also interested in watching what others have to say about the show. This alerts to the use of the volume of tweets as the only reflection of viewers’ social experiences when engaging with the show.

Taken together, these studies indicate that social engagement with certain television content is influenced by two main types of factors: internal, which can be the self-presentation feeling or the sense of social connectedness; and content-drive, the resultants emotional responses, discontentment, or opinions about the television programs.

2.4 Predicting Television Show Success

This section will follow up on the aforementioned and introduce the main techniques used when studying television show success. The success of Television has been traditionally measured through the number of people watching a program, what is called viewership; or TV rating, which is defined as the percentage of total TV households tuned in to a program at a given time given the entire population of TV-equipped homes; or share that is the same percentage but expressed as a percentage of the audience that was actually watching TV at the time. However, given the changes in audience behaviour, these metrics are no longer sufficient, and we must seek new, more accurate metrics (Hill, 2014).

In more recent years, with the boom of social media, the viewers’ social online engagement was taken into consideration when measuring the success of a TV show (Hill, 2014; Shestiyuk et al., 2019). In addition, many studies have relied on the use of social media by analysing its predictive power for the TV audience, and evidence was found that social media data can be used to predict viewership (see, for example, Hill (2014)).

Much research has been contributed to the study of prediction and analysis of television audiences sizes. Numerous models have been applied to these forecasts, mainly at the aggregate level and considering a different type of data in their nature. For example, Weber (2002) stated that both general linear models and Neural Networks produce the most accurate short and long term forecasts for the viewership data of the eight major German TV networks. The Symmetric mean absolute percentage error (SMAPE) errors were between 15 % and 28 %.

Meyer and Hyndman (2006) investigated how forecast aggregation influences the accuracy of predictors on three levels: population, segment, and individual. Their report suggests that regression models slightly outperform decision trees and neural networks on all aggregation levels, and the most accurate forecasts are achieved at the population level. The utilised data were individual viewing and demographic data.

Nikolopoulos, Goodwin, Patelis, and Assimakopoulos (2007) applied Multiple Linear Regression, Simple Bivariate Regression, several Neural Networks, and predictors based on the nearest neighbour analysis and human judgement on the audience ratings of Greek TV and compared the performance in terms of mean absolute error. The highest accuracy of around 9.0 was achieved by using 5-nearest neighbours and simple linear regression.

Linear regression models are commonly chosen in prediction when there are several independent variables, regardless of their accuracy, which could be lower either due to the high complexity these models may include or because they are too simple (Nikolopoulos et al., 2007). A multiple linear regression tries to estimate or predict the relation between a dependent variable - the factor that will be predicted, and independent variables - the factors used to predict the value of the former, through a dataset containing observations for both of these variables.

Through an exhaustive study using hourly viewership time series for the top 100 TV networks in the US, Khryashchev et al. (2019) examined the performance of four forecast models: seasonal averaging as a baseline predictor, Facebook's prophet, Fourier extrapolation, and XGBoost under an ensemble predictor that aims to reduce the negative impact of overfitting. The findings actually showed that an assortment of models combined with an ensemble method results in better accuracy than any single method. The improvement concerning a naive baseline method was around 11 %.

Shestyuk et al. (2019) investigated whether population-level behaviours such as TV viewership and Twitter activity could be predicted using EEG measures by testing two separate step-wise linear regression models. The results revealed that EEG measures explained over 60 % of variance in TV viewership and Twitter activity, explaining a meaningful part of the relationship between the two behaviour measures. Finally, Deitz, Royne, Peasley, Coleman, et al. (2016) revealed the strength and importance of neuro engagement measures as social media engagement predictors.

Ensembles methods are widely used across many fields, including in Television studies (see, in addition to the above, for example, Sereday and Cui (2017)). This technique is known for producing improved results as it uses multiple learning algorithms in a combined way. Random Forest, Decision Trees and XGBoost are the most common and successful machine learning methods of ensembles methods.

Gradient Boosting (Freund, Schapire, & Abe, 1999) is a type of machine learning that generates a prediction model from an ensemble of weak prediction models, *i.e.*, in an iterative way, it combines a strong learner from learners that are slightly better than random. Commonly these ensembles are constructed from decision trees. Falling in the boosting category, the trees are added one at a time in an attempt to correct the errors from the previous trees. The model fitting is usually done using arbitrary differentiable loss functions and a gradient descent optimisation algorithm. In practice, this means that the model is fit as the loss gradient is minimised. The loss function can be defined by the user and depends on the problem to be solved.

In case of the model entirely fits the error residuals for some loss functions, the error residuals become zero in the next iteration, and the process ends prematurely. This may lead to overfitting (Bentéjac, Csörgő, & Martínez-Muñoz, 2021), and regularisation factors may be explored to control the additive process of gradient boosting.

XGBoost (Chen & Guestrin, 2016), for Extreme Gradient Boost, is a decision tree ensemble technique based on Gradient Boosting. The high scalability, speed and efficiency of XGBoost are the factors attributed to its success. In the same way as Gradient Boosting, it builds many predictive models that together offer considerably more accurate predictions than any one model could, and it is done by minimising a loss function.

Figure 2.3 shows an example of how XGBoost would work to predicts if a person likes

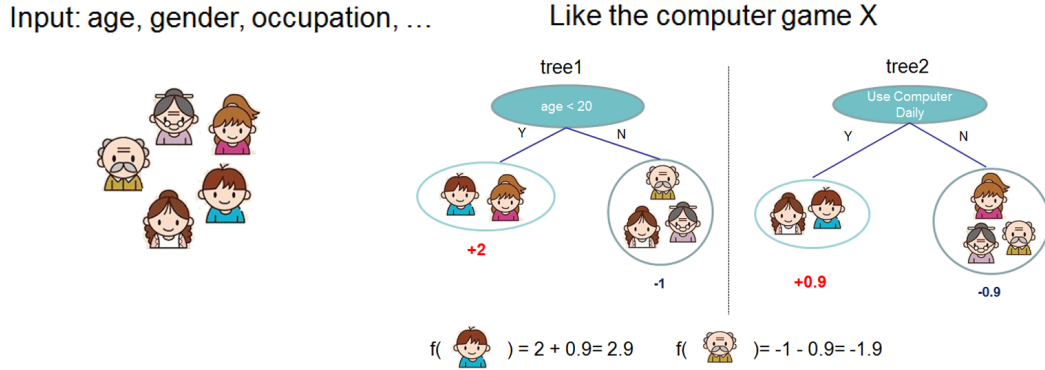


Figure 2.3: Tree Ensemble Model ¹

computer games based on their age and gender. Once it receives an input, it is placed in the trees' root and follows the decision rules till found a leaf node with a prediction. In the end, the predictions from all trees are combined. The combination may be the sum of all predictions or the mean.

In addition, new research for unstructured data, including neuro data and social media data, and new methods such as Artificial Neural Networks (ANN) have been applied to predict television audience ratings and viewers' social online engagement (Venkatraman et al., 2015) due to the overall growth of data availability. For example, Multilayer Perceptron networks (MPL) and Radial Basis Function networks (RBF), which are two kinds of ANN models, were used by Guixeres et al. (2017) to investigate whether brain response, heart rate variability and eye-tracking could be used to predict views for advertising on Youtube. The results confirmed the effectiveness of ANN to accurately classify and predict the number of viewers of each ad via their biometric response, with MPL networks working better than RBF networks. However, it is unknown to what extent we can replicate these results in video content such as television shows.

The Back-propagation Network model was used by Cheng, Wu, Ku, and Chen (2013) for predicting TV audience ratings using various data from Facebook fan pages of TV dramas. This study proved the use of this type of data to perform TV audience ratings as feasible. The Mean Absolute Percentage Error (MAPE) was between 6 % and 24 %.

The presented research shows that Artificial Neural Networks can be used to predict audience behaviour accurately. ANNs models are inspired by the structure and functioning of the nervous system to simulate the learning capacity of the human brain in the acquisition of knowledge. The brain consists of many neurons connected in the sense that when they receive inputs from the sensory organs, they transmit electric signals to each other, which are then combined and processed until they arrive in various output nodes.

The ANN model emulates this biological neural network in its architecture. The first layer contains the set of input factors that are transformed into a set of output factors, the output layer. The hidden layers are responsible for this transformation through a function that depends on weights determined from the set of training examples. The value of neural networks applications

¹Adapted from: <https://xgboost.ai/2016/12/14/GPU-accelerated-xgboost.html>

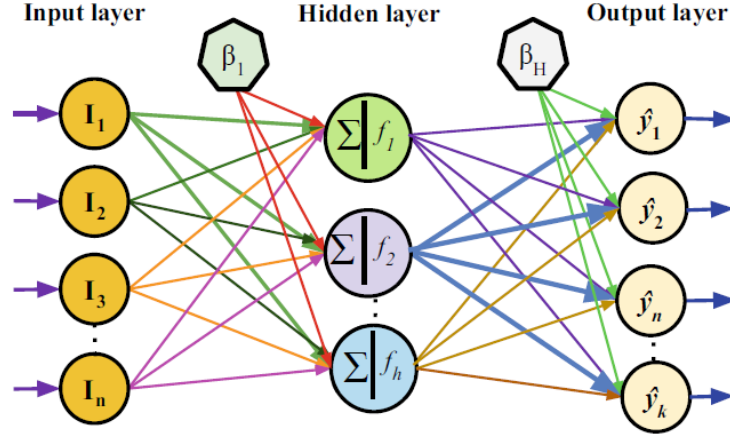


Figure 2.4: Structure of MLP Neural Network.
Retrieved from Heidari et al. (2020)

is given by the significant utility for pattern recognition, classification, learning or analysis of incomplete or noisy data, to name a few. Moreover, these models prove to be a promising alternative to traditional techniques for non-linear temporal processing (Dorffner, 1996).

There are many kinds of neural networks, mainly determined by the definition of their architecture. The feed-forward Multilayer Perceptron network, or MLP for short, offers support for dealing with temporal dimensions and thus can be applied to the prediction of time series. The preparation of the data is a crucial step while using these models. Particularly, the lags that may occur in observations must be flattened into feature vectors. This ANN usually performs univariate modelling of time series by inputting a given number of lagging terms in the time series and projections as outputs. A general structure of an MLP network is shown in Figure 2.4 with one hidden layer. The input layer and output layer are linked by activation functions, the most widely used is the S-shaped sigmoid function (Heidari et al., 2020).

Chapter 3

Case Study

In this chapter, data is presented in detail, its main features and how we access it. The software used to do the analysis and models' implementation is introduced as well. Next, we provide an overview of the methodology taken to solve our problem, the chosen techniques, and the developments. The remainder of the section is dominated by the training phase of the models and their evaluation.

3.1 Business Understanding

Before going any further, it is essential to contextualise the scope of the project and the added value it may bring. As already mentioned, this project is done in collaboration with MindProber. MindProber delivers valuable and comparable engagement metrics by integrating cutting-edge neuroscience, remote sensing, and AI technology to the media industry edged to know the most engaging moments during airtime.

MindProber collects physiological data, namely electrodermal activity, from hundreds of participants wearing psychophysiological sensors during the visualisation of specific television content, such as soap operas, sporting events, live news, political debates, and reality shows.

The remainder of the section will explain in more detail how the data collection process is done, what media content includes, and the desired output.

3.1.1 Procedures & Participants

Participants were recruited through publication on the company's website and social media pages. They were selected to represent the Portuguese population in a wide range of demographics regarding gender, age, and geographical location. As concern to age, the participants were classified according to age groups as follows: 18-24, 25-34, 35-44, 45-54, 55-64, +65. From the moment a participant joins the panel, they receive invitations to participate in the different studies. The participant may accept or refuse to participate in each monitoring session, depending on their availability. That means that each participant may be in more than one monitoring session.

The minute-by-minute Twitter volume data was extracted directly from Twitter for each episode. It represents all tweets posted during air time that contain specific hashtags related to the show. All data is kept anonymous to respect consumer's privacy.

3.1.2 Monitoring Sessions (Stimuli)

After joining the panel, participants receive a biometric sensor at home and the instructions for downloading and installing the app that allows synchronisation with the content.

Participants watched the live television shows at home, in their regular environment, without any behaviour restrictions. This way, it is recreated, as much as possible, a “naturalist” setting in which the participant is watching the TV screen in the same way they watch it without wearing the biometric device. In the meantime, data is synced with ACR. The collection of data is at the same time as the program is airing, so there may be commercial interruptions.

3.1.3 Business Core

MindProber selects TV shows based on their impact expectation, tending to develop own-initiative studies in television debates, premieres of premium time or a cup final game, for example, and also in a perspective of continuity as the case of consecutive episodes of soap operas or entertainment programs. Aside from working on the needs of those who are looking for a way to analyse media content.

The value of this project is especially important on pilot-test, although it is not limited to it. More specially, by collecting galvanic skin responses for episodes that have not been aired yet, the company can forecast the impact of the media content by identifying the most engaging moments, the segments more favourable to lead the viewers to shift channel and anticipate the popularity of the same. This way, the producers can remove those moments that have low engagement, leveraging success.

3.2 Data Analysis

3.2.1 Episode Level Analysis

The original data set includes forty episodes over several television shows, where we may have more than one episode for the same television show, spanning nine different genres. The genre and the respective number of episodes in the sample, indicated in the parentheses, are the following: Car Racing (1), Mini-Series (1), Political Debate (1), Reality Show (8), Reality Singing Competition (1), Soap Opera (1), Sports (24), Talent Show (2) and Talk Show (1). Table 3.1 is gathering elements of Television Show in more detail, some that will be used and explained in later analysis.

Selected television shows are representative of the current landscape that we can find on television by including several types of shows. However, it may be not easy to found comparable characteristics between them since representing so varied content. Data for each episode were collected from 15-69 participants, with an average of 37 participants per episode. The air-time of

Television Show	Show Type	Number of Participants	Average Twitter Volume	Episode Duration
E7	Car Racing	69	12.57	01:45
E8	Mini-Series	47	12.8	01:00
E6	Political Debate	40	12.04	01:49
E4	Reality Show	34	8.63	00:59
E17	Reality Show	34	21.43	01:02
E23	Reality Show	30	34.92	00:59
E26	Reality Show	47	97.86	02:47
E27	Reality Show	49	99.52	03:01
E30	Reality Show	55	156.24	03:05
E31	Reality Show	35	226.94	01:33
E33	Reality Show	41	288.09	01:36
E28	Reality Singing Competition	30	130.02	01:30
E13	Soap Opera	46	16.84	00:57
E1	Sports	15	1.91	02:18
E2	Sports	55	2.59	02:13
E5	Sports	53	10.8	00:59
E9	Sports	51	14.28	02:55
E10	Sports	44	14.89	02:14
E11	Sports	24	15.51	01:21
E14	Sports	58	18.1	01:15
E15	Sports	22	18.14	01:08
E16	Sports	17	19.85	01:07
E18	Sports	24	23.49	01:14
E19	Sports	31	24.11	01:10
E20	Sports	29	24.57	01:11
E21	Sports	21	25.7	01:13
E22	Sports	30	29.6	01:59
E24	Sports	12	67.18	02:26
E29	Sports	31	150.11	01:57
E32	Sports	113	282.55	02:03
E34	Sports	45	327.56	01:58
E35	Sports	49	339.44	01:57
E36	Sports	111	345	01:14
E37	Sports	94	378.38	02:16
E38	Sports	48	498.68	02:13
E39	Sports	47	579.88	02:42
E40	Sports	31	3143.83	04:23
E3	Talent Show	69	4.6	01:03
E12	Talent Show	29	16.64	01:26
E25	Talk show	31	69.21	01:00

Table 3.1: Television Show information

each episode also provides a wide range: the episode with the shortest duration was fifty-seven minutes long, and the longest one was four hours and twenty-three minutes long.

Episode-level analyses were carried out to assess differences in Twitter activity and EDA activity. As the selected shows represent different realities concerning its projection on Twitter - for example, reality shows are more likely to generate social media activity than sports since, in reality shows people are involved in the stories and want to share their opinions, while in Sports people are supporting their teams and therefore more focused on the screen and less open to distractions, we expected to find significant differences among episodes of different genre.

For each episode, numbers of tweets for each minute were averaged across the show duration, i.e., we get the number of tweets per minute. The values range from 1 for episode E1 to 3.143 for episode E40, with a mean equal to 189 and a standard deviation of 502. Thus, the results are indicative of a high variance in Twitter activity among shows. The same can be confirmed visually through the inspection of the boxplots in Appendix A. There were also found differences in Twitter activity across episodes of the same type for reality shows, sports and talent show episodes. The focus on these three types has to do with the fact that we have a more robust sample.

The comparisons also revealed the following pattern on Twitter activity among the different types: Sports have higher social activity than reality shows, and reality shows have more activity than talent shows. Although this might be a surprise, an episode is skewing the results around sports, given a high volume of activity.

Since we found some considerable difference across the different show types and within each type, and since the number of episodes was low for some and did not allow to access confident analysis, the further analyses were performed only on Sports episodes and after removing outliers. This way, it is removed cross-type differences, and we target our attention to one phenomenon only.

The episodes excluded from the analysis within Sports were episodes with low activity, where most points revealed no activity and one episode with many social buzzes that represent a particular event with global impact. In terms of ID, episodes E1, E40 and E2 were not part of the study. By doing so, we are removing atypically situations that may compromise our results, but there is still variance in the data to capture different realities.

3.2.2 Electrodermal Signals

The electrodermal activity was collected using a Biometric sensor that participants wear while watching the content. The resulting data structure is one time series for each participant, comprising a timestamp and corresponding measurements of skin conductance. The time points are obtained on a second-by-second basis in real-time. In this context, the device collects the data that will be used to measure the level of engagement through electrodermal activity, given by the changes in sweat gland activity. As participants can interrupt the viewing in the application whenever they need to, there may be some periods with a lack of data that last until they resume the viewing.

Figure 3.1 displays an example of the raw signal for a particular tester of episode E14. We can see that the time series shows several ups and downs through time with different amplitudes. This is because this participant did not use the sensor for a few seconds, which caused a cut-off in the time series, as shown in the figure.

Skin conductance readings were divided into the phasic and tonic components - throughout a linear and a non-linear transformation of the signal. These transformations were developed through the “ledapy” package in Python. Thus, there are four metrics from the phasic component and two metrics from the tonic component in addition to the raw signal. This phasic-tonic separation allows isolating the parts of the signal related to the events, i.e., the phasic component, from the overall level, i.e., the tonic component. Although both components are important,

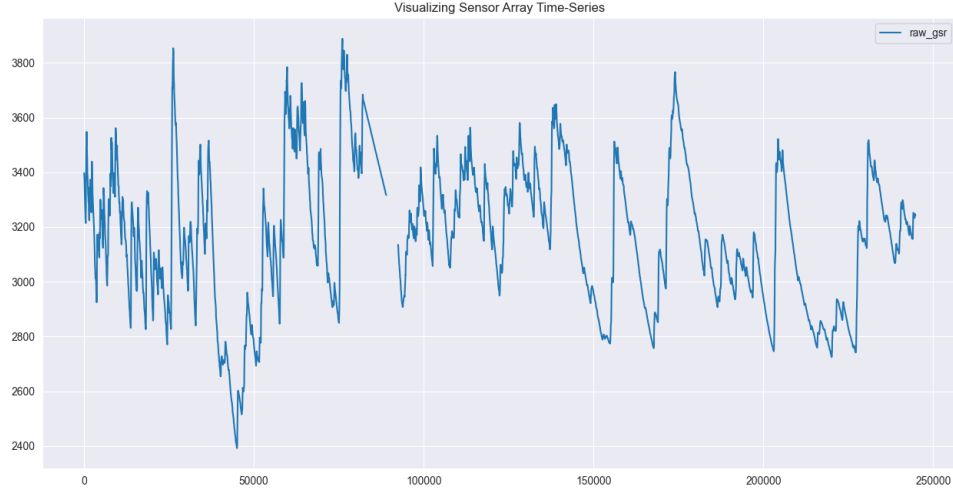


Figure 3.1: Electrodermal Activity across episode duration (Episode E14)

since our interest is to associate moments of higher arousal and engagement at particular times to social engagement, the phasic component is expected to be the most relevant feature of the present study.

Figure 3.2 summarises how the EDA signal was decomposed into new individual EDA measures.

Participant data with excessively contaminated signals, *i.e.*, excessive missing points in the entire episode, were excluded from the analysis. The rate of exclusion was set up on a 10 % allowance for missing values. Given the size of the data sets, it was preferable to lose information than fill it with unrealistic values. For those participants with less than exclusion rate, the values were interpolated through linear interpolation. The missing point is filled with the midpoint between the two points surrounding the missing point in this method.

The success of exclusion was confirmed by the reduction of the number of participants between 2 and 35 participants in each episode.

The analysis of each episode will be performed on clean data with 7 participants, which is the minimum number of participants we have. The choice of which participant to include in the analysis was in a random way. In this way, we preserve the initial construction of data and ensure the representation of all types of behaviours.

As has been said, our goal is to anticipate how an audience will react during media exposure, measuring social engagement. Although arousal, here measured by the skin conductance responses, may depend on multiple antecedents, the stimulus itself is of higher interest in media influence research. If any media influence occurs, we may expect that variation in interpersonal arousal would be somewhat related to the power of the content to generate a response in all audiences. Thus, the chosen parameters to measure arousal should detect similar patterns and reassembles between the different participants. Given the lack of demographic measures, which

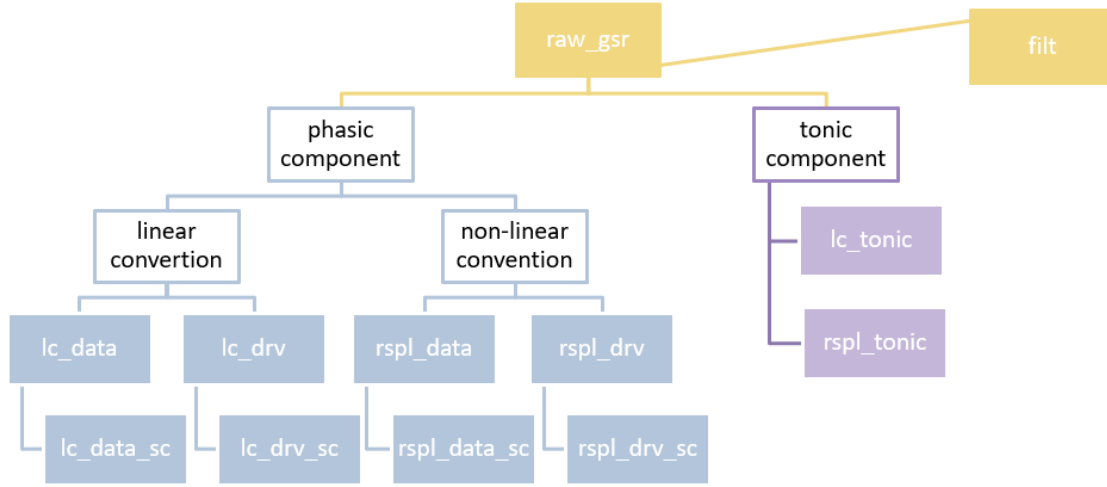


Figure 3.2: Decomposition of the EDA Signal

would give us insights into the profile of each group, correlation and variance were the chosen metrics to provide some indicative estimates about the similarity of the individuals. The correlation is a way to summarise data and gain insights into the relationships between the various participants.

The data set of each participant containing the raw signal of the skin conductance activity were correlated between them using Spearman's rho. This way, we accounted for the non-linear nature of the signal. The final value was calculated as the mean of these correlations. If the correlation between participants is high, then they share similar patterns that are very likely to be stimulus-related. On the other hand, a weak correlation will suggest that other factors besides the stimulus may determine the responses.

The results shown in Table 3.2, in general, indicates that the participants do not present a similar behaviour during air-time. The mean value across episodes was computed and was placed in the last row. The value also sustains the conclusions. Allow we have exceptions, for example episode E21 and E39. To support the analysis, correlation matrix showing the correlations between the participants were obtained and are attached in B. Figure 3.3 shows two of them in which we can confirm the above by observing that for episode E21, the matrix indicates that the majority of the participants are highly correlated. However, for episode E9, the same is not valid, and we may derive that participants do not show similar behaviour.

These findings may have an impact later in the prediction model as the input data may contain contradictory signals, where the signal can be going down and up for the same minute. So the model may not found so quickly the relationship between the input and the output.

Complementary, it was computed the mean and variance between participants for each episode minute that were then aggregated into a plot (available on Appendix C). The analysis of these graphs are in concordance with what was previous derived as, in general, there is a high variance across participants during the episode. It also shows that, in general, as the episode comes to an end, the electrodermal activity tends to increase. However, this may be an excepted result

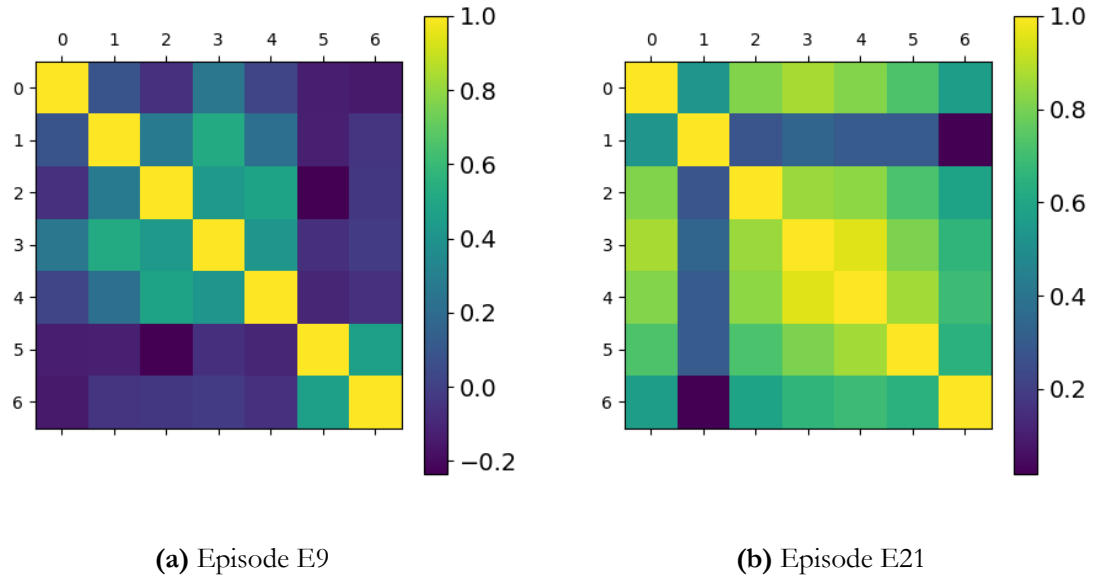


Figure 3.3: Matrix of Spearman's correlation between participants for the raw signal of electrodermal activity

because when the game is almost finishing, people tend to be more involved and excited waiting for the final whistle or a last-minute point. To accomplish this analysis, the raw signals were normalised for reasons that will be later explained.

Data Preparation

Before applying any analysis or measurement, and in a way to obtain reliable results, it is a good practice to ensure that the electrodermal signal does not contain any significant artefact and the signals are in a scale that allows between-participants comparisons.

The magnitude of Skin Conductance Responses can vary a lot from tester to tester, given the nature of the skin conductance. Each body has a personal ability to conduct current and typically correspond to different sweat levels across individuals. Given these variations, the data was corrected to allow for a direct and meaningfully comparison between individuals. Another result of this transformation is that no single participant's measurement will be weighed higher only because its absolute skin conductance values are higher.

The standardisation method used to correct the inter-individual variance was the Proposition of Maximal Response, where each individual SRC data was divided by the participants maximum SCR (Braithwaite, Watson, Jones, & Rowe, 2013). From this, the minimum value is assumed to be zero, and the maximum is the result of an extraordinary stimulus.

Likewise, it is also recommended to transform the Skin Conductance Level into a standardised form. SCL measurements were corrected by applying a Range Corrected Score (Braithwaite et al., 2013) via the Equation (3.1)

Television Show	Mean inter-subject correlation Value
E24	-0.044
E15	-0.041
E22	-0.009
E36	0.056
E10	0.068
E9	0.113
E32	0.127
E5	0.156
E34	0.157
E29	0.163
E19	0.168
E18	0.2
E11	0.254
E38	0.264
E35	0.306
E16	0.321
E14	0.392
E37	0.43
E20	0.451
E39	0.722
E21	0.847
Average	0.243

Table 3.2: Mean inter-subject correlations for the raw signal of electrodermal activity for the whole episode

$$SCLDataPoint = \frac{SCL - SCL_{min}}{SCL_{max} - SCL_{min}} \quad (3.1)$$

The need for normalisation is also supported by the fact that the electrodermal activity might not always follow the normal distribution.

GSR artefacts were isolated and removed from the raw data during preprocessing by applying a high-frequency noise reduction filter developed by MindProber. An artefact is any fluctuation in the signal that skin conductance changes have not caused, for example, derived from abrupt movements or external factors. This smoothed signal was then converted to z-scores across time within each participant. Further analysis will make use of this new variable that represents a pure signal.

To match the temporal resolution of the Twitter activity, the second-by-second points were averaged for each minute of the program within participants. As a direct result, we lose the values for the first minute. This segmentation on a minute basis, although it may include random fluctuations in sweat activity, it was preferred over a more aggressive smoothing as this could leave out many valuable changes, especially in episodes with low average Twitter activity.

Since we are interested in testing to what extent we could use a group metric to predict Twitter activity, a new experimental setup was designed. The second-by-second standardised SCR and SCL values and EDA signal were averaged for each minute of the episode across participants. As such, we only have one value for each variable: the mean value of all participants.

Considering that Twitter activity and Electrodermal activity may fluctuate at different scales and the latter is unable to represent the valence, we hypothesised that the rate of change, *i.e.*, minute-by-minute differences in values, would be a more pertinent metric instead of the raw values alone. In this way, all values were converted to the first discrete differences in both experimental setups.

Nevertheless, in an earlier analysis, this assumption was confronted with the hypothesis of using the number of tweets as a metric that would be further transformed to the rate of change. Since the last did not offer significant improvements, the former was preferred.

3.2.3 Twitter

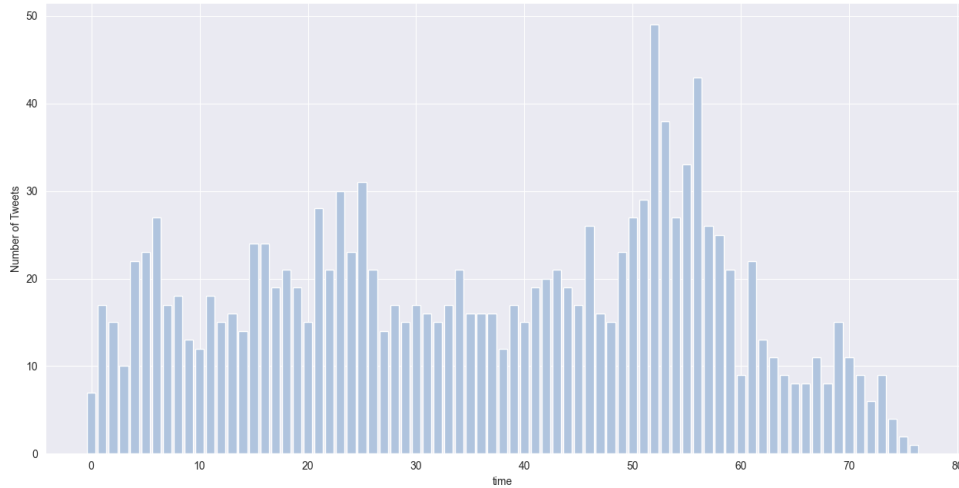


Figure 3.4: Twitter volume across episode duration (Episode E14)

Minute-by-minute Twitter volume data, or the total number of tweets, for each Television show episode, were obtained directly from Twitter. Activity per minute was chosen over activity per second given the high number of moments without tweets' activity, particularly for the shows with less projection. Due to privacy concerns, only anonymised and aggregated Twitter volume values were used, namely the total number of tweets and retweets.

The tweets represent all tweets posted during the broadcast of each episode that includes specific #hashtags related to the program. The reason for choosing only tweets posted during the broadcast time was to have messages that corresponded to the event's content. In other words, we wanted to look for comments that were intended to be posted at that specific time.

The hashtags used are the official hashtags designated by TV shows' networks and communicated to viewers through social media or during emissions. Although we are ensuring that the tweets include contextual information of the shows, as Twitter users are not forced to use hashtags in their posts, we may be excluding relevant tweets.

Due to data protection, there is no information regarding the participants' social media activity, which means that it is not guaranteed that participants are Twitter users.

The Twitter Volume during the air-time of episode E14, taken as an example, is shown graphically in Figure 3.4. It shows that we have segments that led to a more significant variation in the activity, some with high amplitude leading to peaks. For this particular example, from a certain point, as the episode was coming to an end, there was a decrease in the number of tweets.

Data Preparation

To remove between-episode variability ($SD=37.8$) and match the electrodermal analysis procedures, Twitter values within each episode were normalised by converting them to z-scores.

As previously stated, there is a temporal lag between the on-screen scenes and the moment tweets are posted online. This lag is likely to be in the order of seconds to minutes and may depend on the episode and its structure. After some experiments and analysis, we did not find a substantial delay, a consistent pattern, across all participants within each episode that allowed us to access the optimal lag. Therefore, after testing for a lag equal to 0, 1 and 2, we decided not to shift the tweet's differences to accommodate this possible lag.

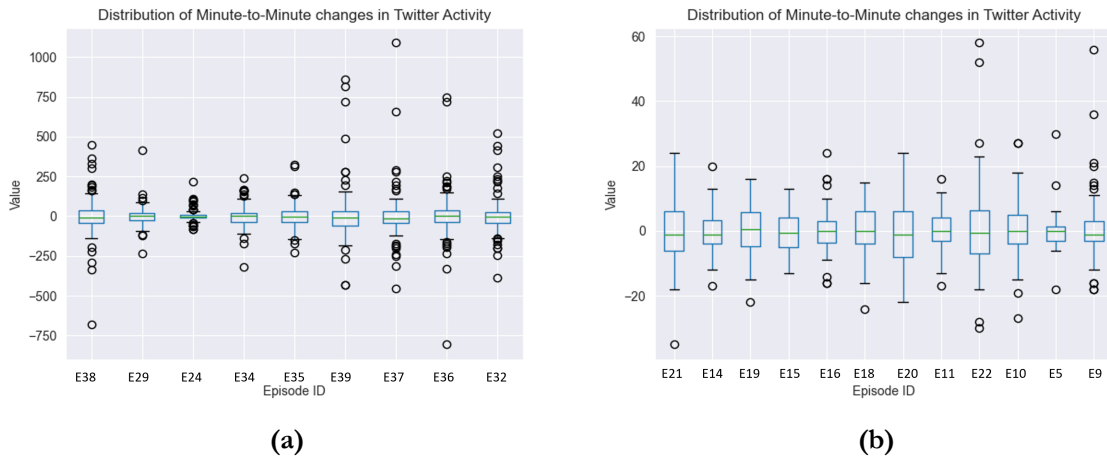


Figure 3.5: Distribution of Minute-to-Minute Differences in Twitter Activity (divide in two groups according to the activity volume)

In the last step, the minute-to-minute changes in z-scored Twitter volume were obtained to match EDA processing. Figure 3.5 shows the distribution of these minute-to-minute changes for each episode. The figure indicates a concentration of the changes around zero, suggesting low activity between minutes; even so, the highest decrease was around 750, and the highest increase was around 1000. This implies the presence of outliers, some observations that may be derived from special show events.

3.2.4 Data Input

Resulting from the previous stage, twenty-one episodes of Sports will be used as the base of work for further analysis and modelling. The variable that we are interested in predict is the changes in Twitter activity from one minute to the following. The attributes available to performer this task are as following:

- Timestamp: minute from the start of the episode
- eda_signal: raw values of electrodermal activity after filtering high frequencies
- lc_data: parameter one from phasic component obtained through a linear conversion of the signal
- lc_drv: parameter two from phasic component obtained through a linear conversion of the signal
- rspl_data: parameter one from phasic component obtained through a non-linear conversion of the signal
- rspl_drv: parameter two from phasic component obtained through a non-linear conversion of the signal
- lc_tonic: parameter one from tonic component obtained through a linear conversion of the signal
- rspl_tonic: parameter two from tonic component obtained through a non-linear conversion of the signal
- tester: participant identifier. Used for prepossessing purposes.

Figure 3.6 shows the relationship between the EDA signal and Twitter Activity represented in minute-to-minute changes in z-score.

In order to give answers to our research questions, several experimental setups refined to the specification of each were set up. Next, we will provide a comprehensive description of these constructed scenarios.

Since for each episode, *(i)* we repeatedly observed the same individuals, *(ii)* we have multiple variables for those individuals, and *(iii)* these observations take place at various points in time, we will be working with panel data. Thus, each data set contains observations for the episode's duration for the seven participants and the correspondent twitter value for that episode.

The first scenario puts in evidence the use of the composite signal as an input variable instead of using the components of the signal as input variables. As a result, we have one data set with the differences in the composite signal for each of the seven participants and the minute-to-minute changes in Twitter activity for each episode minute. When we are using the signal components, the data set is made up of the minute-to-minute differences in each of the six components across the seven participants and, again, the minute-to-minute changes in Twitter activity for each episode minute.

Following, each of these data sets is modified to accommodate new elements. The number of participants will change to investigate the impact of the number of the sample on the performance. All the experiments will be replayed for five and one participant.

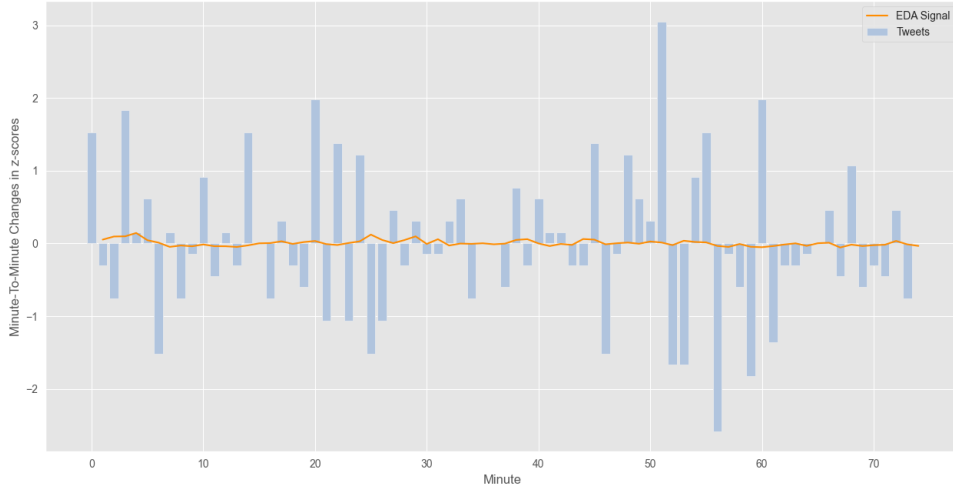


Figure 3.6: Minute-to-Minute Changes in Twitter Activity and EDA Signal in z-scores (Episode E14)

Additionally, we will analyse the impact of using individual data against group data. Here, the group data is given by the mean of all participants. Hence, the data set will change lightly to include only one value for each variable as input.

The characteristics of the analysed data sets are shown in Table 3.3, which displays for each data set the name by which we will identify, if the data set is related to the components or the composite EDA signal, the variables in use, the approach taken, whether it is population-level or individual-level, and, in case of individual level, the number of participants in the analysis.

Name	Variable's Level	Variables	Approach	Nr Participants
C1	Components	lc_data,lc_drv,rspl_data,rspl_drv,lc_tonic,rspl_tonic	Individual-Level	7
C2	Components	lc_data,lc_drv,rspl_data,rspl_drv,lc_tonic,rspl_tonic	Individual-Level	5
C3	Components	lc_data,lc_drv,rspl_data,rspl_drv,lc_tonic,rspl_tonic	Individual-Level	1
C4	Components	lc_data,lc_drv,rspl_data,rspl_drv,lc_tonic,rspl_tonic	Populational-Level	
R1	Composite Signal	eda_signal	Individual-Level	7
R2	Composite Signal	eda_signal	Individual-Level	5
R3	Composite Signal	eda_signal	Individual-Level	1
R4	Composite Signal	eda_signal	Populational-Level	

Table 3.3: Properties of the studied data sets

At this point, we have prepared and accommodated our data to be used by Machine Learning algorithms. The next step is to divide the data into train and test data. Once the model is trained on the training set, the test set will be used to evaluate the model's performance. Since the latter is composed of different observations from the ones in the training set, we will assess whether the model has learned to generalise from the training set on unseen observations or if it has just memorised it.

On that account, we left out one episode that we will use later on in the evaluation part. The episode was chosen randomly, and it corresponds to episode E18. The training data set then contains the remain twenty episodes.

3.2.5 Software

The software chosen to complete all the analysis and implementation was Python, version 3.9. Python is a free programming language widely used in every industry and scientific field, including Data Science and Machine Learning. The simple and easy to learn syntax, the code readability and the dynamic edit-test-debug cycle makes this tool very popular. This software also features some modules and packages previously implemented that encourage code reuse when using such functions in constructing the model.

PyCharm, an open-source application that allows running Python code and Python libraries, will be used as a development environment to support data transformation, data visualisation, and model implementation.

Python is a dominant player in AI branches, such as neural networks and machine learning, with incredible success. However, given the large amounts of data required to process, some of Python’s flaws may pop up, such as speed, dynamic typing, or memory consumption. Nevertheless, for the data sets in use and having in mind the potential of this software in machine learning, this software seems suitable and reliable.

3.3 Modelling

Considering all earlier mentions about the data and the desired output, we reduce our problem to a Supervised Learning problem. Supervised Learning algorithms aim to model relationships and dependencies between the output of the target predictor and the input characteristics so that the output values for new data can be estimated based on certain relationships that have been learned from previous data sets.

Motivated by the increased use of second-screen activity, earlier works investigating Twitter activity, and having the collaboration of MindProber with their domain expertise of TV advertising, we aim to use Artificial Neural Networks and Ensemble Trees to predict Twitter volume across program sections.

One of the biggest challenges when applying and modelling an algorithm for prediction is dealing with data differences. Next, we summarise some particularities that have been mentioned in this work. First, data granularity is a concern in our problem since electrodermal data is stored second by second, and Twitter activity is represented on a minute-by-minute basis. Therefore, the first intuition was to aggregate the participants’ biological data for each minute to match the temporal structure of the other variable. We preferred this approach to other techniques, such as Mixed Data Sampling (MIDAS) regressions, where we could work with different frequencies at a time, as the difference is slight. Using second-by-second data, we could be masking essential variations in the signal.

Besides the above, as it is not sure that there is a linear relationship between the variables and the target, we must seek models that can handle non-linearity and complex data structures.

3.3.1 Modeling Techniques

As aforementioned, our episode observations are in a multi-dimensional data set. For each episode, the participants change; nevertheless, we can find more than once the same tester. This means that we cannot merge our episodes in one data set, as we would be mixing data from different participants. To face this singularity, we propose to build an ensemble predictor that utilises one base model to produce the predictions by episode.

The ensemble was developed in Python by scratch, considering the particularities of our data, and it will work as follows. First, we will be training a base model, the same one, on each episode data set. The resultant predictions will be then aggregated by the median of all predictions, resulting in one predicted value that gives the central tendency of the predictions' distribution. Figure 3.7 shows how we constructed our ensemble through the aggregation of individual model's performance in a single prediction.

Our emphasis will be on comparing Extreme Gradient Boosting and Multilayer Perceptron Neural Network as base models for our ensemble. We also tested other algorithms such as Decision Trees, Random Forest or Stepwise Regression models. Still, the error values were higher than those obtained by the chosen models, so they were not considered in later stages.

What concerns the remainder of the section will describe and analyse the training phase on the experimental setups that were previously mentioned.

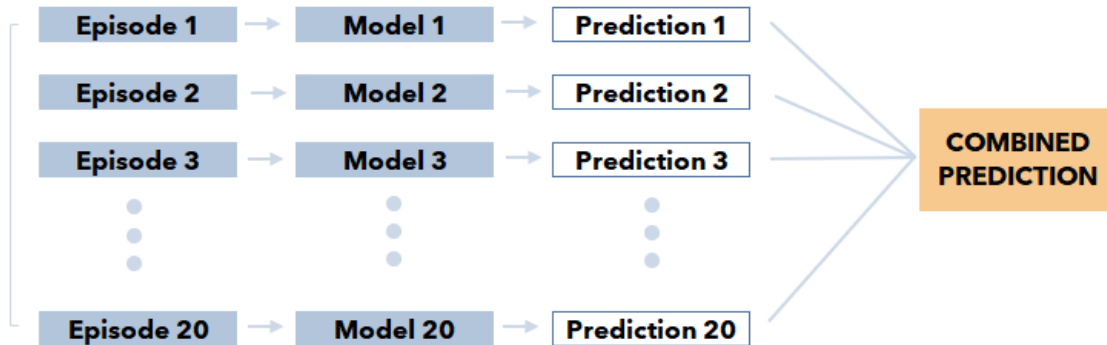


Figure 3.7: Ensemble Architecture

3.3.2 Model Training using EDA Components

We begin with the data set that comprises the phasic and tonic variables data of the seven chosen participants. We trained XGBoost and Multilayer Perceptron Neural Network models for this experiment, and the results will be displayed. Then, in a second stage, it will consider the variants to this data set. More precisely, these variants concern to the modification of the number of participants under consideration and the grouping of individuals into a single metric.

First Framework: Extreme Gradient Boosting

Extreme Gradient Boosting, known as XGBoost, is a decision-tree-based ensemble algorithm that uses a gradient boosting framework. It provides both computational efficiency and high effectiveness, which were the reasons to choose this model: fast to execute and good performance. The best performance is the one that best controls the over-fitting.

XGBoost can be directly applied to regression problems, and it assumes that all variables are numeric, which is guaranteed for our data. Therefore, the first step was the model's configuration, which entails the choice of parameter values. Once achieved, the model will be used to train the datasets.

For these experiments, we used the implementation of the 'scikit-learn' package for XGBoost.

Hyper Parameter Setting

The model training process is controlled by hyperparameters whose values are defined before beginning the learning process and are external to the model. As they cannot be estimated from data, it is convenient to tune them by testing their effect on model performance.

The hyperparameter was adjusted using the Grid Search technique developed from scratch. A dictionary of the hyperparameters to evaluate containing the parameter name and an array of values to try were specified. The output displayed the combination of the hyperparameters and the correspondent value of Mean Absolute Error (MAE) and Mean Squared Error (MSE), which will be adequately explained later. The defined Grid Search process was then designed to train and evaluate one model for each combination of the hyperparameters. Once completed, we analysed and chose the best combination, which gives the most accurate model. We reserve two episodes to evaluate the performance from the training data set, and we used the remaining to do the grid search.

The value range for each parameter was defined after trial and error tests, done to give us sensibility on which zone the best values might be. With this non-exhaustive approach, we overcome the computational power that the process would require if we attempted to evaluate every single combination.

XGBoost comprises three types of parameters: The general parameters are related to the boosting phase. The booster parameters depend on the chosen booster - here, we only worked with the most common and influential of these parameters. The learning task parameters depend on the type of machine learning problem. Given that we have a regression problem, we assign this parameter to regression with squared loss.

The parameters evaluated include the following ones:

- The number of decision trees used in the ensemble is an important hyperparameter for the XGboost. More trees are usually better as the model works sequentially to improve predictions made by previous trees. The argument "n_estimators" controls the number of trees, and it is frequent to set to a large value, although a high number is reflective of more computational power. It takes values between 0 and 1. The search interval was set to [500,1500,2500,3500].

- Each tree is trained on a portion of the training dataset chosen randomly, and its values are between 0 and 1. Lowering this value introduces more variance and allows good generalisation. The “subsample” argument specifies the number of samples used to fit each tree, and it is between 0 and 1. In turn, the argument “Colsample_bytree” rules the number of features used to fit each tree, set at 100 per cent as default. In this case, all features are being used in the training dataset. The range choose for both of these parameters was [0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0].
- The “Max_Depth” controls the tree depth, i.e., the maximum number of child nodes for each tree. It ranges from 0 to 1, and although lower values help stop model complexity, we might lose relevant information present in the data if too small. The default value is six, and the interval of search ranges from 0 to 30.
- The amount of model’s contribution on the ensemble prediction, known as learning rate, is controlled via the “eta” argument. High values may lead to optimum values not being found, while with too low values we might be sub-optimising the values. The default value is 0.3. In our search, we looked through the [.3, .2, .1, .05, .01, .005] range.
- The “min_child_weight” is a threshold that defines the minimal result for the sum of the weights in the child. It ranges from 0 to infinite. The values to search are inside [1, 30].

It is essential to notice that we do not use the completed list of hyperparameters for XGBoost; instead, we optimised the most common used. However, the results depend on the interval we defined, during our investigation and resources, so we cannot guarantee that we achieve the optimal solution for the present problem.

For the final configuration of the model, we propose to use as XGBoost parameters: 10, 5, 0.05, 0.7 and 3500 for max_depth, min_child_weight, eta, col_sample_bytree and n_estimators, respectively.

At last, a seed equal to 32 was defined to make the results reproducible for the same data set at different points in time.

Evaluate the model: k -folder cross validation

When employing machine learning algorithms, a common practice is to do multiples runs or repeated cross-validation and evaluate the performance across the results, especially when the learning algorithm is stochastic. For example, we may think that our model performs well when it may be overfitting the data. Cross-validation is a widely used validation technique that reserve a sample of the data set and use the remaining part to train the model. Then the model is tested on the reserve data set. This way, it is possible to estimate the model’s predictive power when using data not included in the training phase.

From the various models available for performing cross-validation, K -folder cross-validation was the method chosen, and it works as follows. First, it randomly splits the entire data set into k folds. Then, for each of these k -folder, a model is built on $k-1$ folds of the data set, and one is left out to serve for testing the model. This process is repeated until each k -folder has been used as the test set. The errors of each prediction are saved and then averaged to provide a performance

metric of the model. This way, we get an indication of how well the method performs on average. It is also a good practice to add a variance measure to analyse the results, such as standard error or standard deviation. The idea is to anticipate how much the skill of the procedure may vary in practice.

The configuration of the k value must be done having the data sample in use in consideration as a poor choice may mask the model's actual skill. Such as returning models with a high variance - when using higher values of k ; or with a high bias - when using lower values.

Figure 3.8 provides a visualisation of a 5-folder validation when training a model.

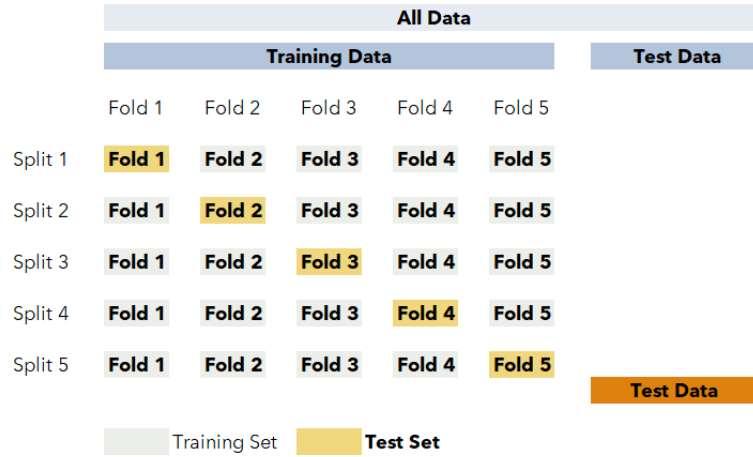


Figure 3.8: Implementation of 5-Folder Cross Validation

For the current project, we will evaluate the model using repeated k -fold cross-validation, through implementation with the 'scikit-learn' package, with five-folds, since our data set is not that large. We will report the averaged Mean Absolute Error (MAE) for each trained model on the ensemble across all repeats and folds and the standard deviation to give a sense of how the model performs on data not used during the training phase. The global mean value is also provided. The MAE is made negative by the 'scikit-learn' package, so it is maximised rather than minimised. To make the interpretation more straightforward, we then multiply the output by -1 to get the positive value.

Mean Absolute Error is a standard metric in regression problems, and it measures the average absolute differences between the actual values and the predictions. The MAE is formally expressed as in Equation 3.2.

$$MAE = \frac{1}{n} \sum_{t=1}^n |(Y_t - \hat{Y}_t)| \quad (3.2)$$

Table 3.4 display the results we obtain, and it shows good scores, with a mean close to 0 and variance values that suggest consistency across the folders, although this may not be so strong for some episodes. The mean and standard deviation across all episodes is presented in the last row. The subsequent analysis will use this tables' structure to display the results. This model will be used later to make predictions on new data.

Individual Data

It was previously mentioned that we are interested in evaluating the prediction power of the model when considering a lowest number of participants. More particular, the two extra experimental setups designed containing data for 5 and 1 participants were used (C2 and C3). The same model was trained using the train data, and the cross-validation results can be accessed in Appendix D. The average mean absolute error of the five folders is not too large; however, the variance is higher than the previous one, suggesting that the model had more difficulty at generalising. The results also indicate that when using only one participant value as input data, the performance is slightly worse.

Population Level Data

In addition to the above, the experimental setup that uses only one point per variable as being the average of all participant's points (C4) were also trained with the configured model. Appendix D comprises the results obtained that seems to be aligned with the other findings: on average, the model appears to be capturing the data characteristics, and the variation across folders is not significant.

Second Framework: Multilayer Perceptron (MLP)

One of the most popular feed-forward neural networks is the Multilayer Perceptron, or simply MLP. The development involves the choice of parameters and the optimiser and loss function involved in the training phase.

The Neural Network Architecture of an MLP is composed of an input layer, a hidden layer and an output layer. This supervised learning algorithm learns a function by training on a data set and discovering the correlation or dependencies that may exist between the inputs and the output in continuous interaction. First, the inputs will be weighted, added a bias, and then passed through an activation function. Next, the given output is used as an input for the following layer, allowing the model to learn from previous mistakes. Finally, the procedure is repeated in a way that error is minimised through an optimiser. Once the process had iterated three times since we have three layers, the forward propagation process ends, and the final output is given.

To train with Neural Networks, data must be numerical, requiring the data to be between 0 and 1. Thus, except for the target variable, which is already scaled, the features were transformed across all data sets using the “MinMaxScaler” from the “Sklearn” package. Then we replay the same transformation on the respective test data set.

For these experiments, the '*MLPClassifier*' regressor from scikit-learn package was used.

Parameter Setting

The optimum parameters for MPL were estimated by tuning the parameters using a grid search as well as using the default parameters of the used packages. We hope to get the most out of the model with this task by finding the best combination of parameters.

On a first trial, the hyperparameters were tuned using a non-exhaustive grid search approach that trained several models on each combination of hyperparameters values from the specified

Study	Mean MAE	Standard Deviation MAE
E21	0.951	0.224
E14	0.925	0.163
E19	0.948	0.127
E15	0.987	0.112
E16	0.865	0.119
E20	0.886	0.109
E11	0.944	0.172
E22	0.920	0.201
E10	0.841	0.123
E5	0.971	0.273
E38	0.789	0.147
E29	0.739	0.179
E24	0.703	0.157
E34	0.884	0.133
E35	0.877	0.132
E9	0.802	0.067
E39	0.643	0.162
E37	0.581	0.151
E36	0.909	0.180
E32	0.788	0.174
Average	0.846	0.155

Table 3.4: Results of the 5-fold cross-validation run: Average and standard deviation of the MAE for XGBoost Model using EDA Components

Study	Mean MAE	Standard Deviation MAE
E21	1.077	0.337
E14	1.307	0.165
E19	1.224	0.316
E15	1.068	0.169
E16	1.165	0.229
E20	1.343	0.196
E11	1.239	0.269
E22	1.735	0.160
E10	1.876	0.287
E5	1.092	0.188
E38	2.034	0.208
E29	1.799	0.158
E24	2.255	0.182
E34	1.793	0.153
E35	1.750	0.231
E9	2.689	0.252
E39	2.492	0.170
E37	1.934	0.293
E36	1.340	0.266
E32	1.948	0.251
Average	1.658	0.224

Table 3.5: Results of the 5-fold cross-validation run: Average and standard deviation of the MAE for MLP Model using EDA Components

space search, scores the performance, and selected the best one. The combination of hyperparameters values that yields the best scores for the considered search is the optimal combination to be included in the model. Next, we present the hyperparameters and the respective search space for the model tuning. The latter were defined after trial and errors tests where we first try intervals with a more considerable step between each value and then narrow it down.

The attributes tested for the neural network, which are the ones that make more sense for the type of problem, were:

- The number of layers in the hidden layer is controlled via “hidden_layer_sizes” argument, and the default value is 100. A more complex task requires a higher number of neurons. The interval search was set to [50, 100, 150, 200].
- The argument ”max_iter” controls how many times each data point will be used, and the

default value is 200. The optimisation loop was defined as [1000, 3000, 5000, 7000, 9000].

- The “learning_rate_init” argument is an indicator of how quickly the model responds to new evidence, and it is a threshold value for the weight updates. We assumed [0.001, 0.007, 0.01, 0.07] as the optimal interval search. The default value is 0.001. This parameter is directly related with “learning_rate” that can assume three forms: (i) it may be ‘constant’ through the process, (ii) ‘invscaling’ meaning that it decrease in a gradually form and (iii) ‘adaptive’ based on the loss.

The activation function in the output layer was left as the default value, which is ‘relu’. Rectified Linear Activation (ReLU) is the most widely used in deep neural networks due to its ability to not activate all neurons at once, making the network effective and sparse.

The Adam optimiser optimises the network, which is also the default value. It uses the square errors as the loss function, which is appropriate for regression problems.

Following the above, the hyperparameters were adjusted in a closed list to control the model optimisation, and the default configuration was also tested. However, since different hyperparameter values can have a different effect in model training, and as our search was limited by computational power and time, it is not ensured that we have found the best possible combination of parameters for the given problem.

On the final configuration of the model, we propose using the default values except for the “max_iter” argument fixed on 9000.

Evaluate the model: k-folder cross validation

Following the same procure, for MLP, an ensemble was trained on the multiple training sets, and its performance is evaluated using 5-fold cross-validation. Once again, the aim is to build a generalised model that can accurately predict unseen data.

Table 3.5 displays the average Mean Absolute Error (MAE) and the standard deviation of the trained model across all repeats and folds. The results indicate a possible discrepancy between the training phase and the future performance of the model when predicting data not used on the training phase if we analyse the standard deviation. However, the average values of MAE are not significantly high.

From table 3.5, we can also see that the model performs very similarly on average across the multiple train sets - the standard deviation for all episodes is equal to 0.224.

Individual Data

To further analyse our hypothesis questions, and as we have been doing, we carry out the same procedure on the data sets with data for 5 participants (C2), data set with data for 1 participant (C3), and data set with the aggregated data from the initial 7 participants (C4). The MLP model, with the defined parameters, was trained on each data set. The average performance of the ensemble model was estimated using 5-folder cross-validation, and the results are summarised in Appendix D.

The model seems to have more difficulty predicting new data from the data set with 5 participants (C2) than in the data set with only one participant. Nevertheless, the standard deviation

across the multiple train models suggests that the performance may vary in practice when applied to unseen data.

Population Level Data

Similarly to the above procedures, we also train an ensemble on data set C4, which comprises the variable's values as the mean of the participants (Appendix D). Again, the results are pretty similar to what we have seen - the mean absolute error is not excessive, although there is evidence of skills deviations in practice.

3.3.3 Model Training using the composite EDA signal

The second stage of modelling will work on the data sets that use the filter value of the composite signal as a variable, either in the one that contains the data for the seven chosen participants as the ones that allow for the modification of the latter in terms of the number of participants and the aggregation of the data. More specifically, we will be using R1, R2, R3 and R4. For each data set, the procedure carried out on the last segment will be replicated, and the results will be summarised so we can efficiently compare both.

The optimum parameters for each model were found by exploring a wide range of possible values using the grid search technique as well as using the default values of the corresponding packages. For each of the two ensembles, these values, the search intervals and default values are visible in Table 3.6 and Table 3.7.

Parameter	Grid Search Values	Default Value	Final Values
n_estimators	[500,1500,2500,3500].	100	4500
Colsample_bytree	[0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0]	1	0.7
Max_Depth"	[0,30]	6	22
eta	[.3, .2, .1, .05, .01, .005]	0.3	0.3
min_child_weight"	[1,30]	1	18

Table 3.6: Default Values, grid search interval and the chosen values for XGBoost Model

Parameter	Grid Search Values	Default Value	Final Values
hidden_layer_sizes	[50,100,150,200]	100	100
learning_rate_init	[0.0001, 0.0007, 0.001, 0.007]	0.001	0.0001
max_iter	[1000,3000,5000,7000,9000]	200	9000

Table 3.7: Default Values, grid search interval and the chosen values for MLP Model

Next, each model was trained using the parameters extracted from the hyperparameter tuning phase using the multiple train data sets. Finally, the performance of the trained models when used to make predictions on new data was estimated with a 5-folder cross-validation. The results for the data set with the seven participants (R1) are shown in Table 4.11 and Table 4.12 for

XGBoost and MLP respectively. The results for the remaining data sets are available in Appendix E following the line that we have been using.

From Table 4.11 and 4.12, we can see that the results are quite similar on average for each model. The mean absolute error is not very close to zero, but it is not significant across folders as well across the trained models. Although MLP performs slightly worse on average across the different trained models, the same is not true when considering R3 and R2, which shows MAE close to zero. For the remaining adapted data sets, no surprises on the results.

Study	Mean MAE	Standard Deviation MAE
E21	1.274	0.280
E14	1.121	0.139
E19	1.245	0.214
E15	1.481	0.277
E16	1.330	0.186
E20	1.223	0.167
E11	1.180	0.188
E22	1.050	0.141
E10	1.096	0.125
E5	0.884	0.243
E38	0.926	0.142
E29	0.880	0.123
E24	0.845	0.084
E34	1.039	0.109
E35	1.005	0.215
E9	0.932	0.112
E39	0.752	0.152
E37	0.820	0.135
E36	1.178	0.167
E32	1.010	0.123
Average	1.063	0.166

Table 3.8: Results of the 5-fold cross-validation run: Average and standard deviation of the MAE for XGBoost Model using EDA Signal

Study	Mean MAE	Standard Deviation MAE
E21	1.315	0.389
E14	1.395	0.22
E19	1.272	0.171
E15	1.318	0.21
E16	1.272	0.214
E20	1.354	0.204
E11	1.394	0.33
E22	1.891	0.253
E10	2.051	0.387
E5	1.156	0.253
E38	2.286	0.259
E29	1.984	0.196
E24	2.383	0.225
E34	2.012	0.200
E35	1.937	0.261
E9	2.905	0.265
E39	2.567	0.162
E37	2.219	0.317
E36	1.445	0.202
E32	2.073	0.272
Average	1.811	0.249

Table 3.9: Results of the 5-fold cross-validation run: Average and standard deviation of the MAE for MLP Model using EDA Signal

3.3.4 Models Assessment

This section described the training phase and the techniques adopted to choose the best models for the data set in use. In brief, the training phase was performed with two ensembles

with different base models - XGBoost and MLP. For each of these models, (i) the optimum set of parameters were estimated using a grid search or the default values and (ii) these parameters were used to train the ensembles on the eight different data sets.

In general, the results were not very distinct across the multiple data sets, which suggests that the model may have captured the characteristics of the data and the relationships between the variables. Although the differences are not significant and change from data set to data set, we may anticipate that XGBoost will perform better on unseen data.

Feature Importance

The decomposition of the EDA signal into different individual EDA scores was designed to assess the importance of each component on the final prediction. Since the phasic and the tonic phases measure different plans of arousal. We hypothesised that the former would be of more value as they are related to the relations to the events. In order to examine the veracity of this hypothesis, we perform feature importance. Feature importance methods help to optimise the model, and they exist to give insights on how the features in the model contribute to the predictions.

The feature importance of each of the variables for each of the trained models was obtained through a built-in method in “sklearn” library for XGBoost, given that it provides easy and reliable results. The final scores were obtained by averaging these contributions. Figure 3.9 shows the final structure of the features in terms of importance to the predictions. We can see that the most important feature according to the build-in algorithm is the “*lc_data*” that comes from the phasic component. The second most important feature is the “*rspl_data*” that also comes from the phasic component, but through a non-linear transformation as opposed to the first that comes from a linear transformation. These features are somehow related, and the importance score is very similar. Then, in third place, we found “*lc_tonic*” that is part of the tonic component, and also “timestamp”. Nevertheless, the feature importance has not shown that there are variables that dominate the model. The distribution is quite balanced.

Even that on the top 3 we have two variables from the phasic component and one from the tonic component, the differences between the two groups are not that significant.

3.4 Evaluation

After the constructed models are used to predict the desired output, the next step is to analyse its results and evaluate the generalisation performance of the ensembles with the chosen parameters. This procedure is a way to know how well our models behave and their adequacy in this problem. The choice of the best techniques to evaluate the models may depend on the type of model and the level of aggregation. It is also essential to choose a metric that supports the comparison across the models. Finally, the evaluation of the model will be done with several regression metrics to allow for a better understanding of the performance.

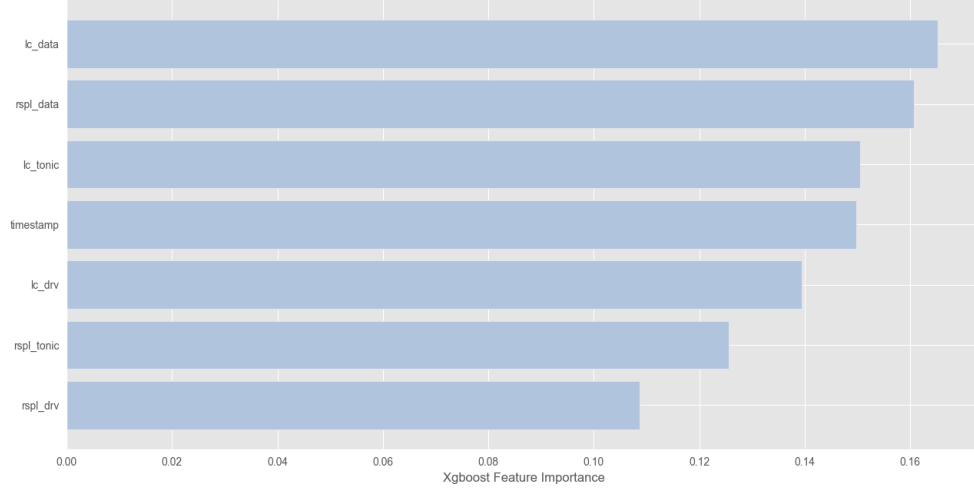


Figure 3.9: Feature Importance for XGBoost

3.4.1 Performance Metrics

Before move on to the results, we will present the chosen metrics with a brief description and the corresponding calculations. That said, the predicted changes on Twitter activity obtained for episode E18 by both ensembles models will be evaluated using MSE and MAE, which was already explained in previous sections and obtained from the Equation 3.2.

The Mean Square of Error (MSE), expressed by Equation (3.3), works with the squared differences between the actual value (Y_t) and the predicted value ($\hat{Y}_t$). In Equation (3.3), the quadratic sum of the prediction error is divided by the number of observations. Thus, it gives more weight to more significant differences. Consequently, the smaller the value, the better is the model performance.

$$MSE = \frac{1}{n} \sum_{t=1}^n (Y_t - \hat{Y}_t)^2 \quad (3.3)$$

Baseline Model

It is common to confront the model results under evaluation against the results of some defined baseline model. The baseline model is usually the simplest possible prediction, working as a threshold that we aim to beat. To do so, we use the bellow relative error metrics that, as they are unitless, we can compare the different setup experiments. The values are expected to be between 0 and 1, with values above one meaning performances as bad as the baseline model.

In light of the problem formulation, we defined the constant model as our baseline. Thus, all predictions will be given by the average target value calculated in the training data, *i.e.*, by the averaged minute-to-minute changes on Twitter activity across the show duration. The predictions will then be combined through the median to meet the resolution used in the training phase.

Even though it uses a pretty base formulation, this baseline is sophisticated as it is an ensemble. The final value is the combination of each individual prediction that may be more realistic and with less variance from the simpler baseline models.

The Normalised Mean Absolute Error (NMAE), also denoted by Coefficient of Variation, is formally defined by Eq.(3.4). The Normalised Mean Squared Error (NMSE) is given by Equation (3.5).

$$NMAE = \frac{\sum_{t=1}^n |\hat{Y}_t - Y_t|}{\sum_{t=1}^n |\bar{Y} - Y_t|} \quad (3.4)$$

$$NMSE = \frac{\sum_{t=1}^n (\hat{Y}_t - Y_t)^2}{\sum_{t=1}^n (\bar{Y} - Y_t)^2} \quad (3.5)$$

3.4.2 Results and Conclusions

	XGBoost				MLP			
Data Set	MAE	MSE	NMSE	NMAE	MAE	MSE	NMSE	NMAE
Baseline	5.80	52.56	1.00	1.00	5.80	52.56	1.00	1.00
R1	9.32	143.32	12.85	2.46	12.49	241.99	125.74	4.83
R2	8.62	104.15	74.29	3.35	6.22	63.55	9.93	1.77
R3	9.04	135.39	233.48	4.52	5.62	50.86	3.19	1.19
R4	7.78	100.72	65.29	3.12	5.48	49.41	2.18	1.10
C1	7.17	84.94	9.17	1.86	22.87	783.25	591.63	10.20
C2	7.28	87.33	6.90	1.77	19.92	604.86	368.27	7.74
C3	8.40	114.47	21.55	2.54	10.17	152.44	87.00	4.34
C4	8.50	103.82	49.56	3.33	10.09	155.30	85.36	4.29
Average	7.99	102.96	52.67	2.66	10.96	239.35	141.58	4.05

Table 3.10: Prediction Errors for XGBoost and MLP

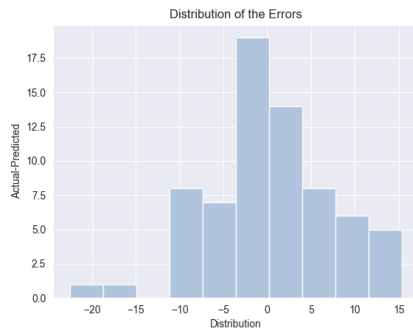
In this division, we do a comparative analysis of the efficiency of XGBoost and MLP models across the eight different data sets that were considered during this project. The performance is carried out in terms of the metrics explained in the previous subsection. The results are shown in Table 3.10 for both models. The results must be read in the following way: each cell indicates the value for each of the metrics for each model and dataset. Cells with the lowest value, not considering baseline, are highlighted in a blue background. The conclusions are related to the models in comparison and for the data sets utilised.

From this table, we can see that the most favourable model, with the lowest error rates, in general, is the XGBoost (in 5 out of 8 data sets). However, it is when we use MLP on R4 - the data set with the averaged signal as a whole for the seven participants, that we get the lowest error values leaving out the baseline.

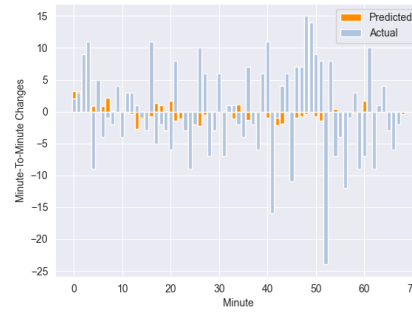
A result that is quite surprising is that the baseline shows better performance than the trained models. The baseline model shows the smallest values both for the MAE and MSE metrics,

although for MAE the differences are not that significant, especially for XGBoost. Furthermore, the normalised versions of MAE and MSE evidence the same. All values are higher than 1, some very pronounced, which means that our model performs worse than the baseline model.

One of the reasons for this behaviour is referent to the way MSE and MAE treat outliers values. The MSE metric, due to its formulation, penalises more severely larger errors. So, if our model makes a very unreal prediction, the squaring part of the function will amplify the error. Even though MAE weighs all errors on the same scale, by taking the absolute values, when we have outliers, the large error will end up being weighted the same way as the lower errors. The direct consequence is that we may have a model that is performing well most of the time but making extreme wrong predictions in some cases.



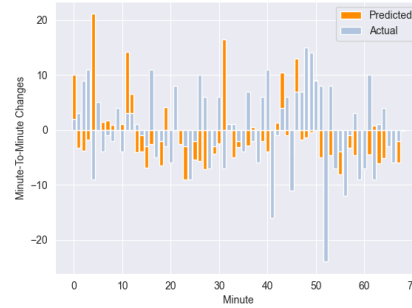
(a) Histograms of Error for MLP on R4



(b) Actual and Predicted values for MLP on R4



(c) Histograms of Error for XGBoost on C2



(d) Actual and Predicted values for XGBoost on C2

Figure 3.10: Prediction of Twitter Activity

To find out if the above applies to our test set, the distribution of the errors, given by actual minus predicted, from the models on the testing set were plotted on a histogram. In general, we would expect to obtain a normal distribution centred at 0, meaning that the magnitude of the error is small. Additionally, the changes across episode duration for predicted and actual were plotted to gain insights on whether the model captures the general tendency or not. All the representations can be found in Appendix F and Appendix G.

An example of the results obtained is shown in Figure 3.10 for XGBoost and MLP on R4 and C2, respectively. As it can be observed, the histograms are not very similar to a normal

distribution. In the MLP model, most of the errors are in the range $[-10,10]$, suggesting that the predictions are more or less constant. However, we have errors outside of this range, indicating higher deviations that may be biasing the metrics. The Minute-by-minute changes in Twitter volume plotted against changes in Twitter volume predicted allows seeing more clear that we have some peaks on the actual twitter activity that are very different from the other points. In general, the model does not seem to be capturing the amplitude of the changes, predicting an almost constant behaviour. However, we may notice that the tendency is not that mismatched, especially for XGBoost.

For the XGBoost, the distributions of the errors show a worse performance, with more unrealistic predictions. In this case, the models seem to overestimate the changes in Twitter activity, and for those points identified as outliers, the difference is even higher than MLP. Nevertheless, on the contrary, XGBoost shows a more dynamic prediction behaviour.

As it can be observed, the performance of the ensembles on the multiple data sets shows some differences. XGBoost appears to achieve better results on the data sets concerning the component's variables, especially for C1 and C2. In contrast, the MLP model seems to be capturing the data relations more easily on the data sets with the composite signal, more specific for R2, R3 and R4. Although the performance on R1 indicates a different direction. This means that the choice between using the signal or its components may depend on the model we are using. The signal as a whole may hold a more complex relationship to the final output, which makes neural networks more appropriate. However, if we have data in a more departing form, XGBoost may give the results needed.

The findings did not reveal a substantial difference in the use of individual data versus the aggregated data of all participants. Even though, as previously stated, the ensemble using MLP with the aggregated metric was the best result comparing with the baseline. Thus, a possible way to improve the exactness of the models when dealing with population data is to use other metrics besides the mean, such as variance metrics that can reflect how different the participants may be.

In addition, it is also unclear the effect of reducing the samples of participants in use as input data. While MLP suggests that one individual is sufficient to discover the characteristics of the data as C3 and R3 shows lower error comparing to C1 and C2, and R1 and R2, respectively, the XGBoost does not show a relevant difference across the data sets.

Despite the above, it seems clear that a more exhaustive parameter optimisation is necessary to further improve the performance of the ensembles and perhaps a more personalised model to adapt it to the characteristics of each specific data set.

A conclusion that may be acceptable is that the changes in the number of tweets per minute may be affected by other factors that are not considered for our models, such as the audience volume or a turning-point on the episode. These factors may be meaningful for the peaks in Twitter activity.

Chapter 4

Conclusions

Nowadays, watching TV and being connected to the Internet is well established as a daily routine for modern society. As a result, show creators, producers and media executives started to include in their strategies this use of social media to attract audience participation and increase viewership. However, producing great shows is also highly connected with the strength of creative content - viewers want to watch emotionally evocative shows. In this study, biologically based responses are integrated to predict the Twitter activity for a set of episodes of different shows. Specifically, from monitoring viewers behaviour while watching a television program and measures their electrodermal activity, and by collecting the total tweets posted about the show during air time, we aim to identify the moments that lead to maximum engagement and those that lead to audience churn through changes in social engagement. That is, by predicting how the engagement is measured by the variation in the number of tweets during airtime.

This project is developed in collaboration with MindProber, a Portuguese startup that provides media content analysis tools energised by neuroscience expertise. The success of this implementation can bring a new way of delivering valuable analysis to their clients about the media content and the way people experience it, working as a bridge to future projects.

MindProber then collected the available data, and it comprises the following features. The physiological data expressed by the sweat gland activity is provided in a time series for each participant on a second-by-second basis. It represents real-time reactions to events on the content. The Twitter activity is given by the total number of tweets in time points but on a minute-by-minute basis. It is an aggregated measure.

The different granularity existent in data- seconds versus minutes, and the differences in data aggregation- individual versus population, makes data preparation and pre-processing a crucial step. Machine Learning algorithms aim to respond to this complexity and provide an improved alternative to earlier works that are mainly focused on traditional methods.

We considered using an ensemble model to predict the outcome by using a base model on the different training data sets to solve this problem. The ensemble models then aggregated the prediction of each base model by the median resulting in one final prediction for the unseen data. XGboost and MLP were the base models used on the developed ensemble. First, the base models' choice was to allow a comparison point between a simpler algorithm versus a more complex one and then based on their characteristics.

Each model was trained on the same training data, which comprises 20 episodes. Before fitting the final model, we attempted to determine the right combination of the hyperparameters that would allow for maximum performance. This process was done using a non-exhaustive grid search or using the default values proposed by the packages used. In the case of XGBoost, the best combination was found through the grid search, as it proved to be more effective. However, contrary to the XGBoost, the default setting was in general preferred for the MLP model. Finally, the performance was evaluated using Mean Absolute Error (MAE), a prevalent metric on regression problems, which measures the difference between two continuous variables, the actual and the predicted.

Once the model's build was accomplished, and the results were satisfactory, the real model's performance was tested on a test set that was intended left out from training. Both models were used in the developed ensemble, and the results were communicated through the MAE, MSE, and the normalised version of these.

At the beginning of our project, we propose to answer a set of research questions that would lead to valuable insights into the social engagement field. First, we did not find statistical evidence that the decomposition of the EDA signal brings more predictive power than the composite EDA signal. Instead, the results indicate similar performance when using the signal as per se and when using its phasic and tonic components. Additionally, it was not proved that the phasic components that are directly related to specific events on TV had more predictive power over the tonic components.

Besides that, the number of participants used for data modelling was under examination. Although the results are not strong enough to conclude, we observed a good performance on the models that used data from only one participant. Perhaps a more focus on the reasons behind this behaviour can be under analysis in future works.

Furthermore, it was also explored the contribution of using a population-level metric as the aggregation of all participant's points in use. Once again, the results did not show a clear position, although they point out some potential in using this approach.

Our final conviction is that the EDA Activity cannot per se completely predict the Twitter activity as other factors directly affect social engagement. For example, the number of people watching a television show directly affects the magnitude of online posting. Therefore, it seems logical that the more people are watching, the more the volume of tweets will be. Another critical factor is the type of episode. For our case, as we are dealing with Sports shows, turning points such as the expulsion of a player or a team scoring a goal may affect more people starting watching. Alternatively, once the result is unlikely to change, people stop watching it. These factors have an inherent influence on social engagement.

Having said that, we suggested the following future work improvements. First, the inclusion of audience or a prediction of the audience as a variable, so the volume of Twitter activity can be easily estimated. If the television show in the analysis is a series of consecutive episodes, it may be good to consider the last audience episode as a variable. An alternative would be the input of the number of tweets minutes before the show goes to air.

We also highlighted the potential of including other variables related to the episode structure, such as adding breaks and relevant happenings. In addition, the level of engagement of the participants could be classified into different levels that could be used as an input variable in the

model. This classification could be derived, for example, by signification reductions of interest or too many variations on the EDA signal.

Another opportunity for improvement may be in a personalised segmentation per scene. People tend to watch the entire act, so providing the aggregation according to it instead of minute by minute may be more accurate in terms of viewer engagement.

4.1 Limitations

During the experiment process, we detected some factors that may impact the findings of our study and are worth noting.

Considering that we are trying to predict a population behaviour from individual data, some characteristics should have been present in our participants. For example, we did not guarantee that the participants were active members of Twitter or other social media that work with posts. This may lead to unrealistic scenarios where the considered participants may not have the instinct to comment and post their opinion while watching television.

Additionally, as the participants are not the same across the episodes, the input data can have different behavioural patterns that can lead to mixed and contradictory signals that the model cannot distinguish as being different.

The fact that the reality we are trying to predict is given for a more varied and vast audience than the one we have, and as the number of tweets is not per participant, we may need more information to generalise our data.

Since participants are not obligated to wear the sensor the whole time, this leads to data quality issues where we have many participants with interruptions. Having a clean signal for most of the episode duration would allow to include more participants in the study.

Finally, and although it was considered a point to improve, the audience size is an essential factor when predicting the changes in the number of tweets. As it was showed on the data preparation step, the episodes in the study represent different shows in terms of success - we have shows with a considerable volume of activity; however, we also have shows with not so significant activity. Naturally, this affects the ability of the model as the expected volume is unknown. As we are only considering the participants' reactions and, although they give us valuable insights, the engagement of the participants within a show is independent of the number of people watching it.

To our best knowledge, this will be the first study using EDA measurements to predict TV viewership and Twitter activity by further exploring the predictive power of Artificial Neural Networks in social engagement prediction. Although the complexity of the problem, we believe that the purpose of this project was overall accomplished. The implementation of the previously referred suggestions and the consideration of the flaws mentioned above will further elevate the study of social engagement through physiological data.

Bibliography

- Arvidsson, A., & Bonini, T. (2015). Valuing audience passions: From smythe to tarde. *European Journal of Cultural Studies*, 18(2), 158–173.
- Auverset, L. A., & Billings, A. C. (2016). Relationships between social tv and enjoyment: A content analysis of the walking dead’s story sync experience. *Social Media+ Society*, 2(3), 2056305116662170.
- Bentéjac, C., Csörgő, A., & Martínez-Muñoz, G. (2021). A comparative analysis of gradient boosting algorithms. *Artificial Intelligence Review*, 54(3), 1937–1967.
- Braithwaite, J. J., Watson, D. G., Jones, R., & Rowe, M. (2013). A guide for analysing electrodermal activity (eda) & skin conductance responses (scrs) for psychological experiments. *Psychophysiology*, 49(1), 1017–1034.
- Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785–794).
- Cheng, Y., Wu, C., Ku, T., & Chen, G. (2013). A predicting model of tv audience rating based on the facebook. In *2013 international conference on social computing* (p. 1034-1037). doi: 10.1109/SocialCom.2013.167
- Christopoulos, G., Uy, M., & Yap, W. J. (2016, 12). The body and the brain: Measuring skin conductance responses to understand the emotional experience. *Organizational Research Methods*, 22. doi: 10.1177/1094428116681073
- Critchley, H. (2002, 05). Electrodermal responses: What happens in the brain. *The Neuroscientist : a review journal bringing neurobiology, neurology and psychiatry*, 8, 132-42.
- Deitz, G. D., Royne, M. B., Peasley, M. C., Coleman, J. T., et al. (2016). Eeg-based measures versus panel ratings: Predicting social media-based behavioral response to super bowl ads. *Journal of Advertising Research*, 56(2), 217–227.
- Dorffner, G. (1996). Neural networks for time series processing. In *Neural network world*.
- Doughty, M., Rowland, D., & Lawson, S. (2012, 07). Who is on your sofa?: Tv audience communities and second screening social networks. *EuroiTV’12 - Proceedings of the 10th European Conference on Interactive TV and Video*. doi: 10.1145/2325616.2325635
- Freund, Y., Schapire, R., & Abe, N. (1999). A short introduction to boosting. *Journal-Japanese Society For Artificial Intelligence*, 14(771-780), 1612.
- Fukushima, Y., Yamasaki, T., & Aizawa, K. (2016, 04). Audience ratings prediction of tv dramas based on the cast and their popularity. In (p. 279-286). doi: 10.1109/BigMM.2016.24
- Gregersen, A., Langkjær, B., Heiselberg, L., & Wieland, J. (2017, 08). Following the viewers: Investigating television drama engagement through skin conductance measurements. *Poetics*,

64. doi: 10.1016/j.poetic.2017.06.002
- Guixeres, J., Bigné, E., Ausín Azofra, J. M., Alcañiz Raya, M., Colomer Granero, A., Fuentes Hurtado, F., & Naranjo Ornedo, V. (2017). Consumer neuroscience-based metrics predict recall, liking and viewing rates in online advertising. *Frontiers in psychology*, 8, 1808.
- Hayes, D. (2015, Jan). *Don't confuse a tv show's twitter impact with its ratings power*. Forbes Magazine. Retrieved from <https://www.forbes.com/sites/dadehayes/2014/11/30/dont-confuse-a-tv-shows-twitter-impact-with-its-ratings-power/>
- Heidari, A. A., Faris, H., Mirjalili, S., Aljarah, I., & Mafarja, M. (2020). Ant lion optimizer: theory, literature review, and application in multi-layer perceptron neural networks. *Nature-Inspired Optimizers*, 23–46.
- Hernandez, J., Zicheng Liu, Hulten, G., DeBarr, D., Krum, K., & Zhang, Z. (2013). Measuring the engagement level of tv viewers. In *2013 10th ieee international conference and workshops on automatic face and gesture recognition (fg)* (p. 1-7). doi: 10.1109/FG.2013.6553742
- Hill, S. (2014, 06). Tv audience measurement with big data. *Big Data*, 2, 76-86. doi: 10.1089/big.2014.0012
- Inc., T. (2011, May). *Watching together: Twitter and tv*. Retrieved from https://blog.twitter.com/en_us/a/2011/watching-together-twitter-and-tv
- Khryashchev, D., Papiu, A., Xuan, J., Dinica, O., Hubert, K., & Vo, H. (2019). Who watches what: Forecasting viewership for the top 100 tv networks. In *International conference on computational data and social networks* (pp. 163–174).
- Kosterich, A. (2016, 04). Reconfiguring the “hits”: The new portrait of television program success in an era of big data. *International Journal on Media Management*, 18, 1-16. doi: 10.1080/14241277.2016.1166431
- Leiner, D., Fahr, A., & Früh, H. (2012, 12). Eda positive change: A simple algorithm for electrodermal activity to measure general audience arousal during media exposure. *Communication Methods and Measures*, 6, 237-250. doi: 10.1080/19312458.2012.732627
- Mantzari, E., Lekakos, G., & Vrechopoulos, A. (2008). Social tv: introducing virtual socialization in the tv experience. In *Proceedings of the 1st international conference on designing interactive user experiences for tv and video* (pp. 81–84).
- Meyer, D., & Hyndman, R. (2006, 11). The accuracy of television network rating forecasts: The effects of data aggregation and alternative models. *Model Assisted Statistics and Applications*, 1, 147-155. doi: 10.3233/MAS-2006-1303
- Moreira, P. S., Chaves, P., Dias, N., Costa, P., & Almeida, P. R. (2018). Emotional processing and the autonomic nervous system: a comprehensive meta-analytic investigation.
- Napoli, P., & Kosterich, A. (2015, 08). Reconfiguring the audience commodity: The institutionalization of social tv analytics as market information regime. *Television & New Media*, 17. doi: 10.1177/1527476415597480
- Nielsen. (2013, Jun). *New nielsen research indicates two-way causal influence between twitter activity and tv viewership*. Retrieved from <https://www.nielsen.com/us/en/press-releases/2013/new-nielsen-research-indicates-two-way-causal-influence-between-/>
- Nielsen. (2014, Apr). *Building time-shifted audiences: Does social tv play a role?* Retrieved from <https://www.nielsen.com/us/en/insights/article/2014/building-time-shifted-audiences-does-social-tv-play-a-role/>

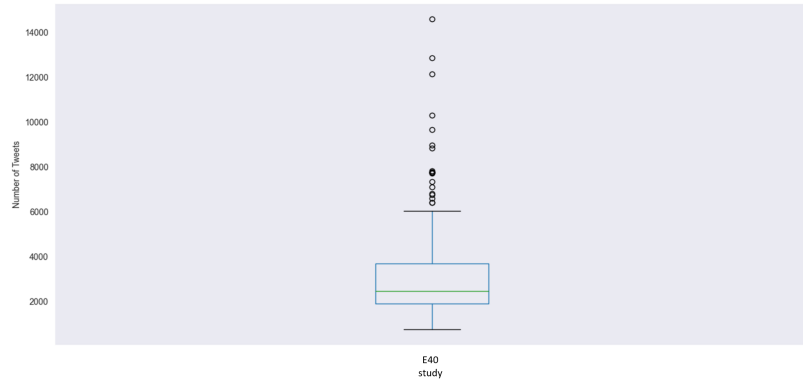
- Nikolopoulos, K., Goodwin, P., Patelis, A., & Assimakopoulos, V. (2007). Forecasting with cue information: A comparison of multiple regression with alternative forecasting approaches. *European Journal of Operational Research*, 180(1), 354–368.
- Patelis, A., Metaxiotis, K., Nikolopoulos, K., & Assimakopoulos, V. (2003, 06). Fortv: Decision support system for forecasting television viewership. *Journal of Computer Information Systems*, XXXXIII, Number 4, pages 100 – 107.
- Peacock, J., Purvis, S., & Hazlett, R. (2011, 12). Which broadcast medium better drives engagement? measuring the powers of radio and television with electromyography and skin-conductance measurements. *Journal of Advertising Research*, 51, 578. doi: 10.2501/JAR-51-4-578-585
- Poels, K., & Dewitte, S. (2006, 03). How to capture the heart? reviewing 20 years of emotion measurement in advertising. *Katholieke Universiteit Leuven, Open Access publications from Katholieke Universiteit Leuven*, 46. doi: 10.2139/ssrn.944401
- Posada-Quintero, H. F., & Chon, K. H. (2020). Innovations in electrodermal activity data collection and signal processing: A systematic review. *Sensors*, 20(2), 479.
- Posner, J., Russell, J. A., & Peterson, B. S. (2005). The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. *Development and psychopathology*, 17(3), 715.
- Pynta, P., Seixas, S., Nield, G., & Hier, J. (2014, 03). The power of social television: Can social media build viewer engagement? a new approach to brain imaging of viewer immersion. *Journal of Advertising Research*, 54, 71. doi: 10.2501/JAR-54-1-071-080
- Richardson, D., Griffin, N., Stephenson, A., Yan, J., Curry, T., Noble, R., ... Devlin, J. (2020, 07). Engagement in video and audio narratives: contrasting self-report and physiological measures. *Scientific Reports*, 10. doi: 10.1038/s41598-020-68253-2
- Schirra, S., Sun, H., & Bentley, F. (2014, 04). Together alone: Motivations for live-tweeting a television series. *Conference on Human Factors in Computing Systems - Proceedings*. doi: 10.1145/2556288.2557070
- Sequeira, H., Hot, P., Silvert, L., & Delplanque, S. (2009, 01). Electrical autonomic correlates of emotion. *International Journal of Psychophysiology*, 71, 50-56. doi: 10.1016/j.ijpsycho.2008.07.009
- Sereday, S., & Cui, J. (2017). Using machine learning to predict future tv ratings. *Nielsen Journal of Measurement*, 1(3), 3–12.
- Shestyuk, A. Y., Kasinathan, K., Karapoondinott, V., Knight, R. T., & Gurumoorthy, R. (2019, March). Individual EEG measures of attention, memory, and motivation predict population level TV viewership and Twitter engagement. *PLoS ONE*, 14(3), e0214507. doi: 10.1371/journal.pone.0214507
- Vecchiato, G., Astolfi, L., De Vico Fallani, F., Cincotti, F., Mattia, D., Salinari, S., ... Babiloni, F. (2010, 06). Changes in brain activity during the observation of tv commercials by using eeg, gsr and hr measurements. *Brain topography*, 23, 165-79. doi: 10.1007/s10548-009-0127-0
- Venkatraman, V., Dimoka, A., Pavlou, P. A., Vo, K., Hampton, W., Bollinger, B., ... Winer, R. S. (2015). Predicting advertising success beyond traditional measures: New insights from neurophysiological methods and market response modeling. *Journal of Marketing Research*, 52(4), 436-452. Retrieved from <https://doi.org/10.1509/jmr.13.0593> doi: 10.1509/

jmr.13.0593

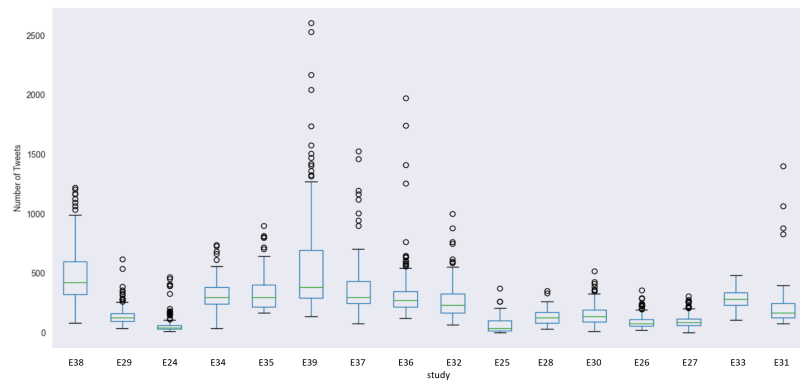
- Wakamiya, S., Lee, R., & Sumiya, K. (2011). Towards better tv viewing rates: exploiting crowd's media life logs over twitter for tv rating. In *Proceedings of the 5th international conference on ubiquitous information management and communication* (pp. 1–10).
- Weber, R. (2002). Methods to forecast television viewing patterns for target audiences. *Communication Research in Europe and Abroad Challenges of the First Decade*. Berlin: DeGruyter.
- Wohn, D. Y., & Na, E.-K. (2011). Tweeting about tv: Sharing television viewing experiences via social media message streams. *First Monday*.

Appendix

A Distribution of Twitter Activity per episode

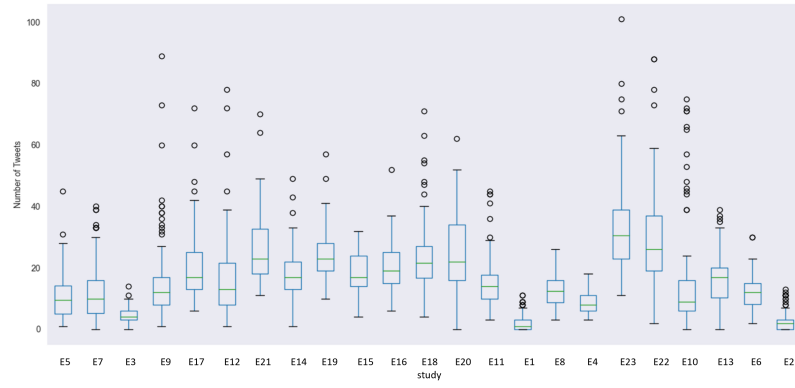


(a) Episode E40



(b) Episodes: E28,E29,E30,E31,E32,E33,E34,E35,E36, E37,E38,E39,E26,E27, E24,E25

Figure 4.1: Distribution of Twitter Activity grouped by volume of tweets



(c) Episode: E1,E2,E3,E4,E5,E6,E7,E8,E9,E10,E11,E12,E13,E14,E15,E16,E17,E18,E19,E20,E21,E22,E23

Figure 4.1: Distribution of Twitter Activity grouped by activity

B Inter-subject Correlations

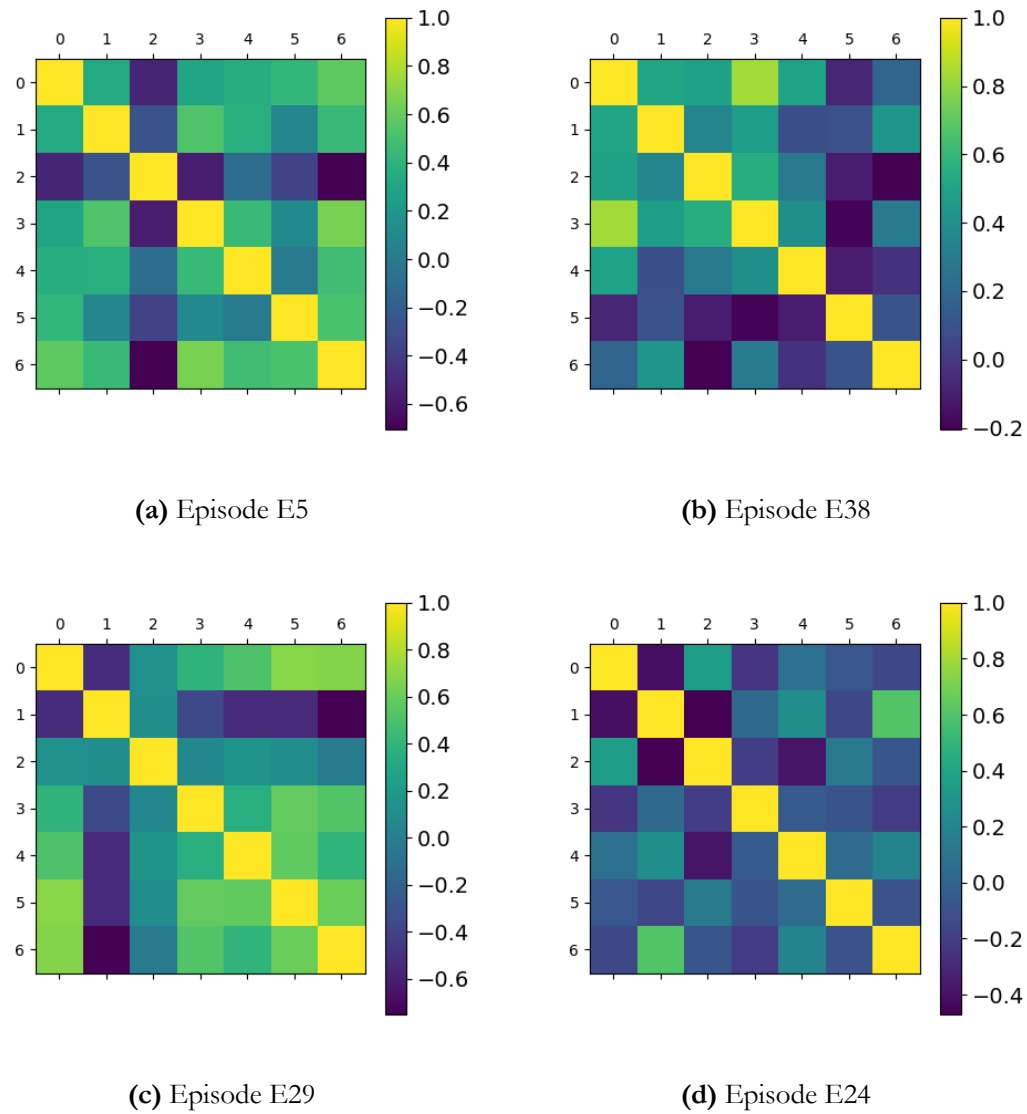
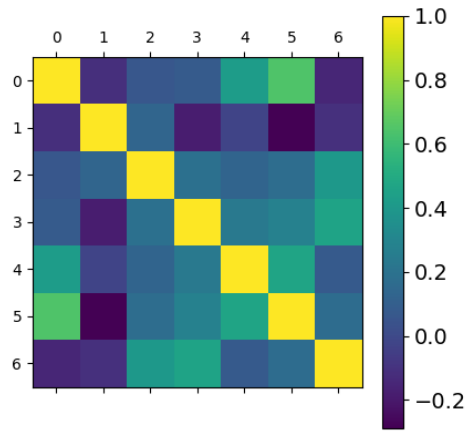
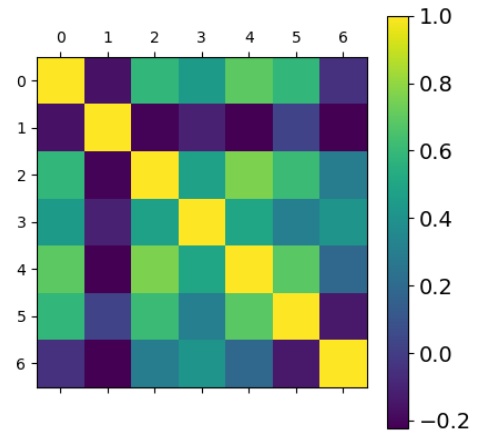


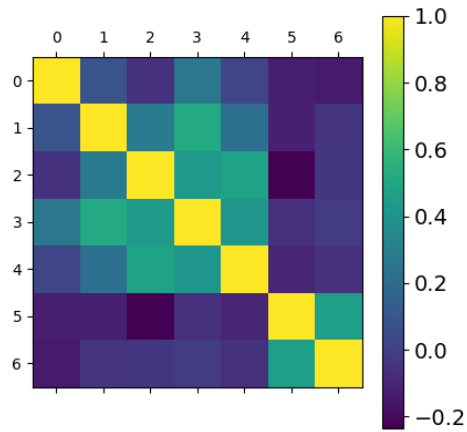
Figure 4.2: Matrix of Spearman's correlation between participants per episode



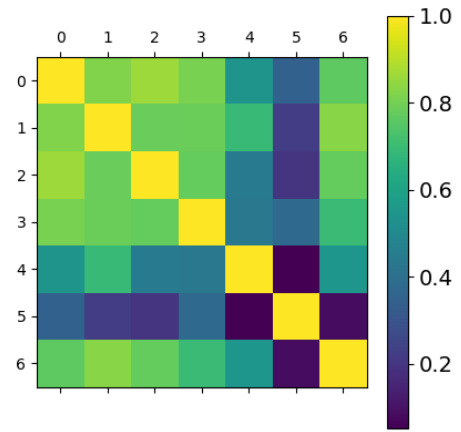
(e) Episode E34



(f) Episode E35

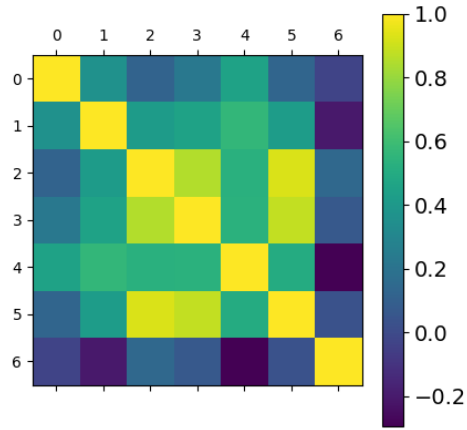


(g) Episode E9

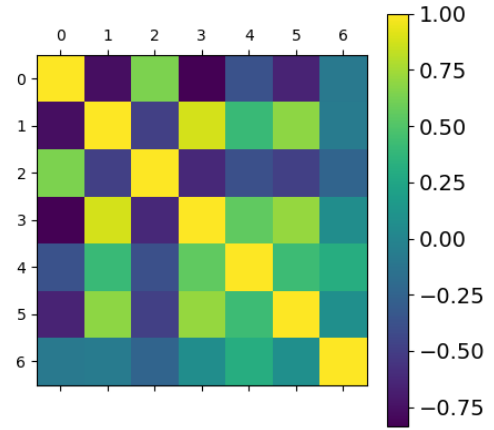


(h) Episode E39

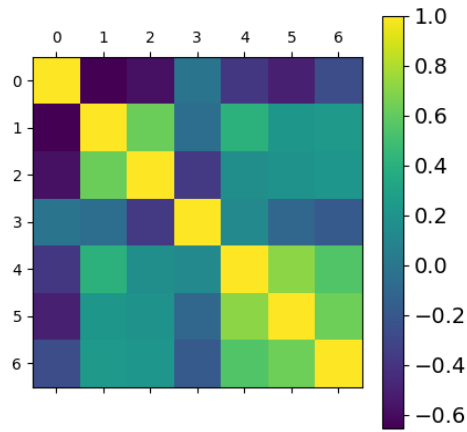
Figure 4.2: Matrix of Spearmans's correlation between participants per episode



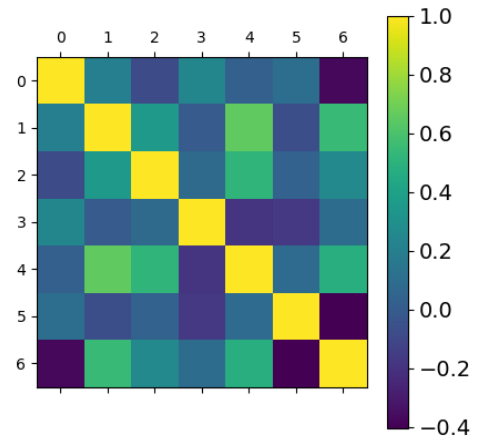
(i) Episode E37



(j) Episode E22

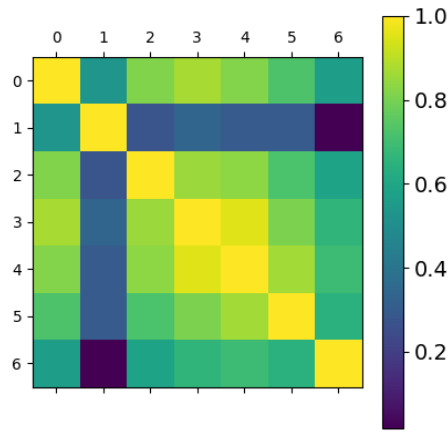


(k) Episode E36

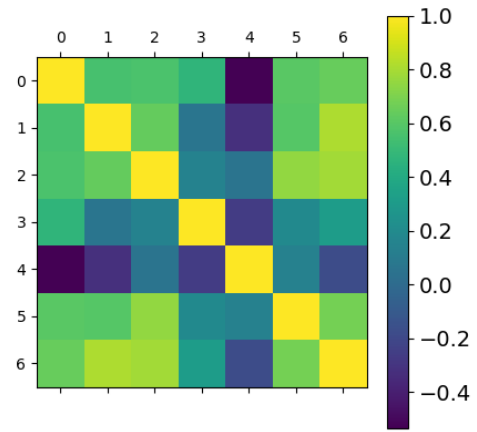


(l) Episode E32

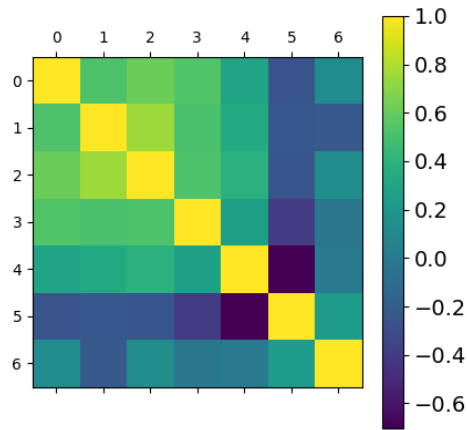
Figure 4.2: Matrix of Spearmans's correlation between participants per episode



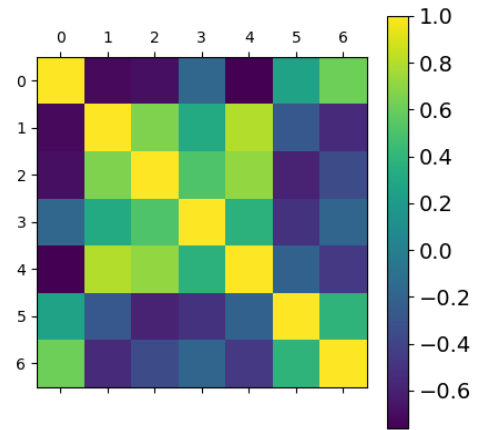
(m) Episode E21



(n) Episode E14

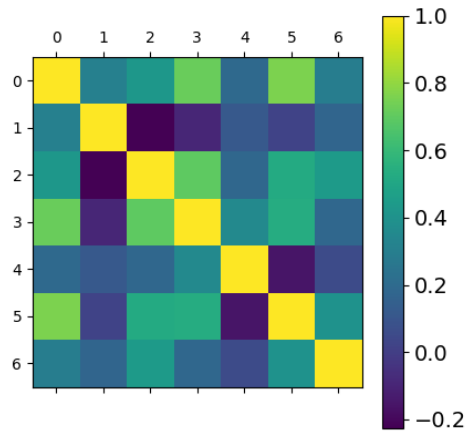


(o) Episode E19

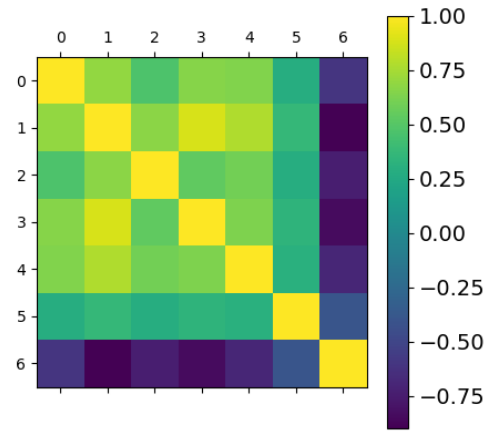


(p) Episode E15

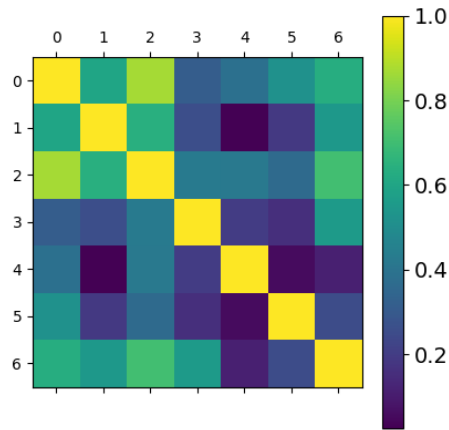
Figure 4.2: Matrix of Spearmans's correlation between participants per episode



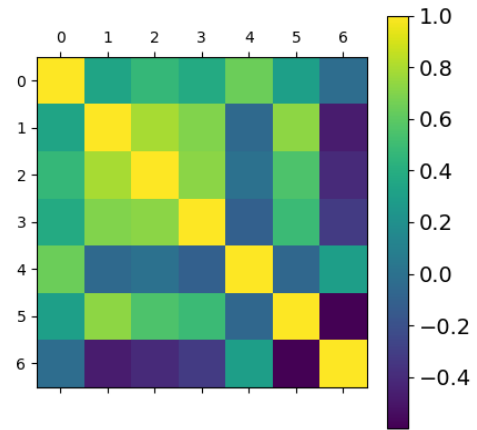
(q) Episode E16



(r) Episode E18

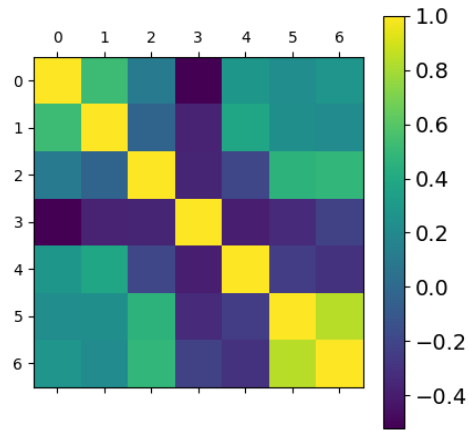


(s) Episode E20



(t) Episode E11

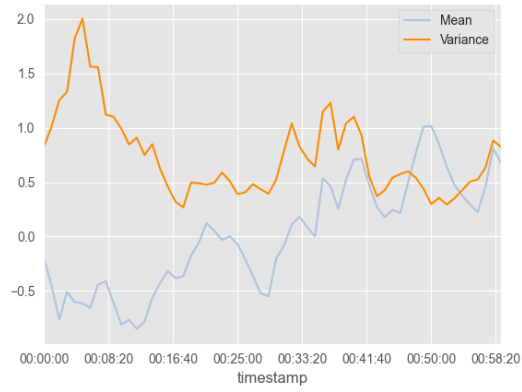
Figure 4.2: Matrix of Spearmans's correlation between participants per episode



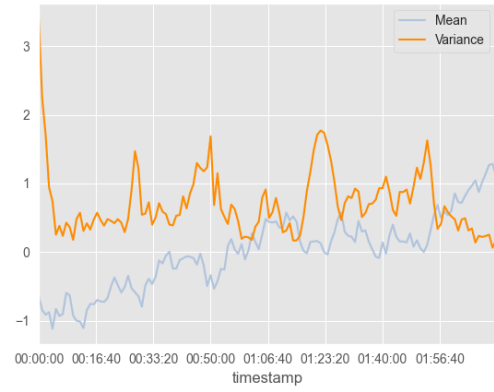
(u) Episode E10

Figure 4.2: Matrix of Spearman's correlation between participants per episode

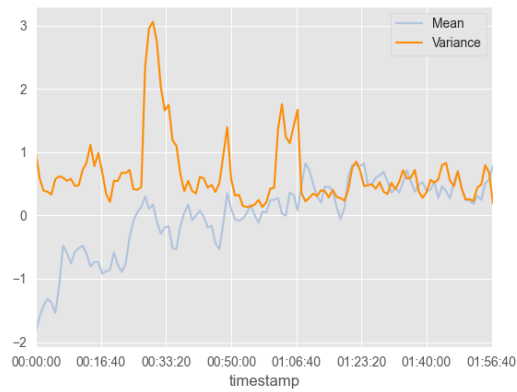
C Mean and Standard Deviation between participants for airtime



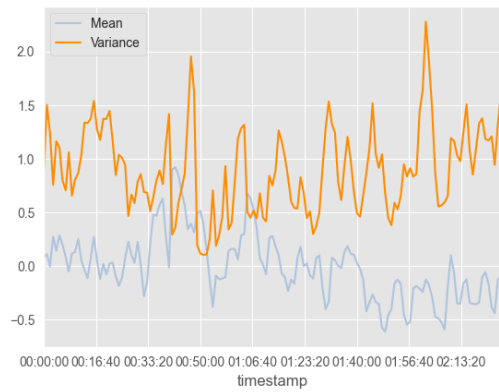
(a) Episode E5



(b) Episode E38

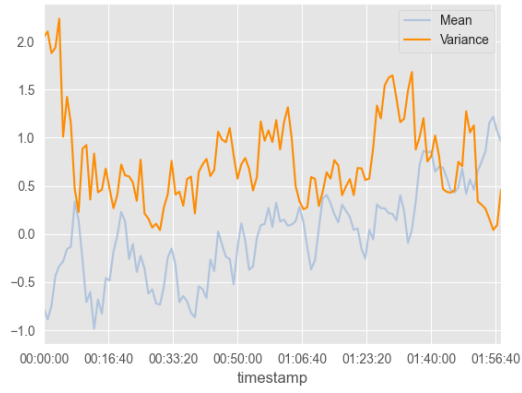


(c) Episode E29

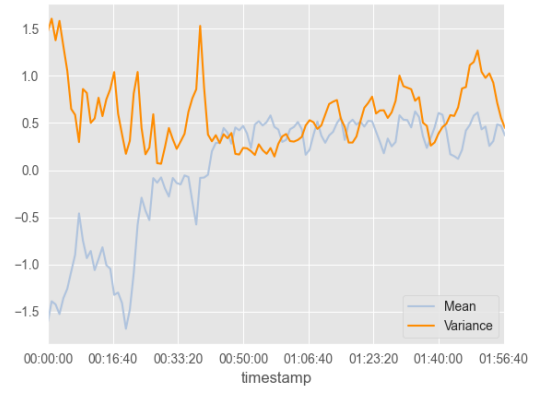


(d) Episode E24

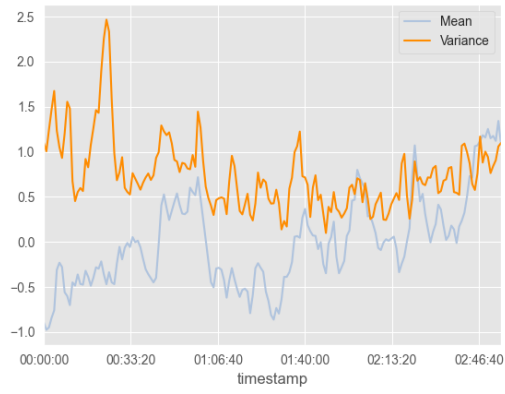
Figure 4.3: Mean and Standard Deviation between participants for airtime



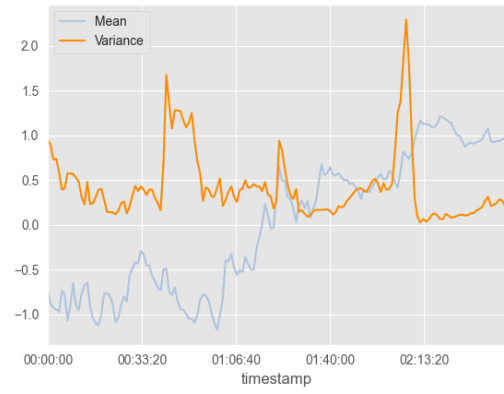
(e) Episode E34



(f) Episode E35

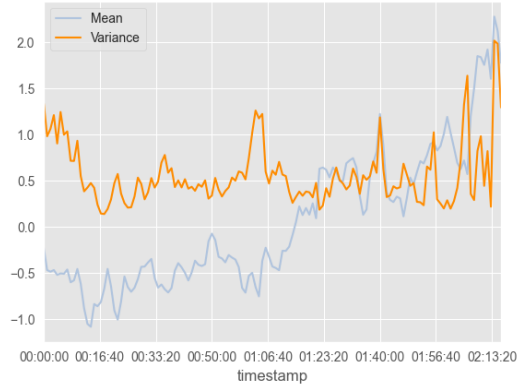


(g) Episode E9

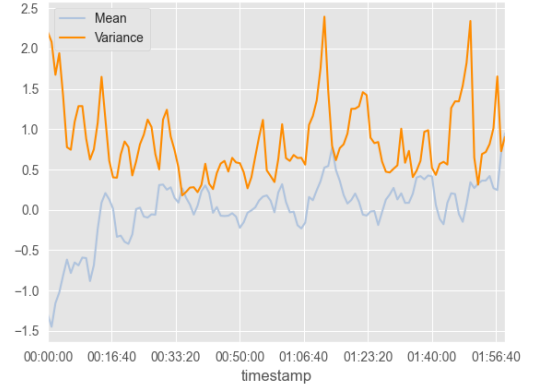


(h) Episode E39

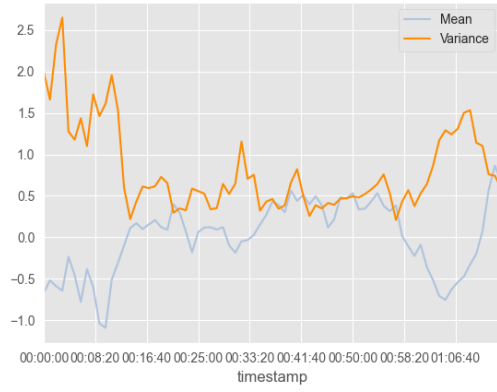
Figure 4.3: Mean and Standard Deviation between participants for airtime



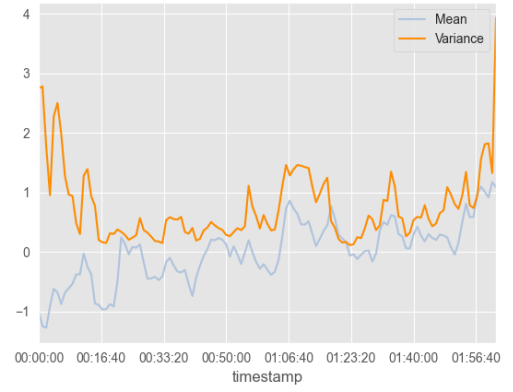
(i) Episode E37



(j) Episode E22

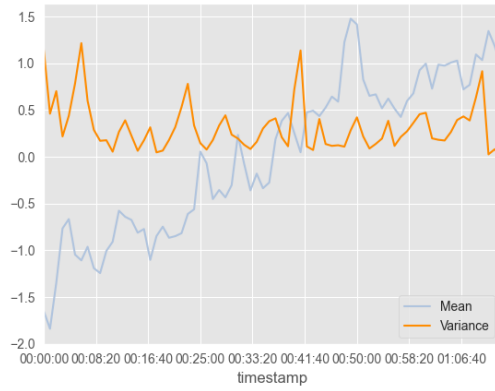


(k) Episode E36

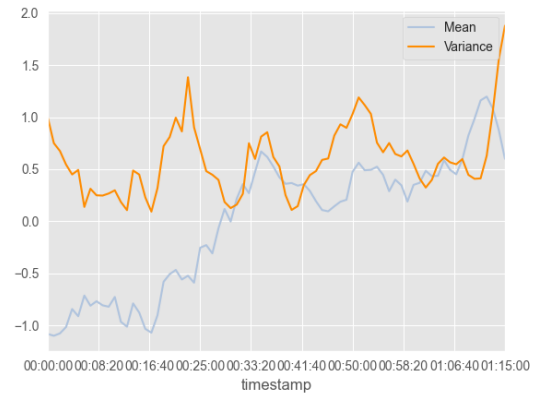


(l) Episode E32

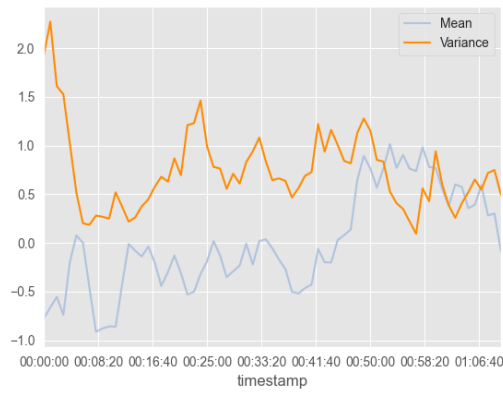
Figure 4.3: Mean and Standard Deviation between participants for airtime



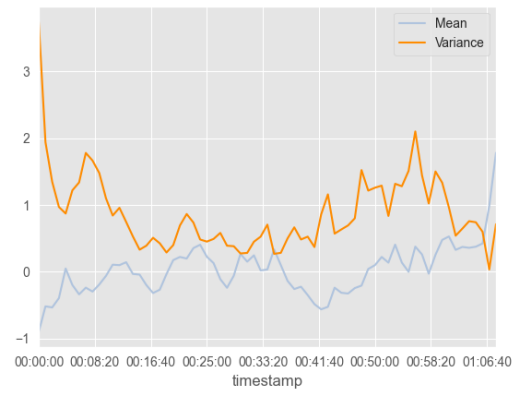
(m) Episode E21



(n) Episode E14

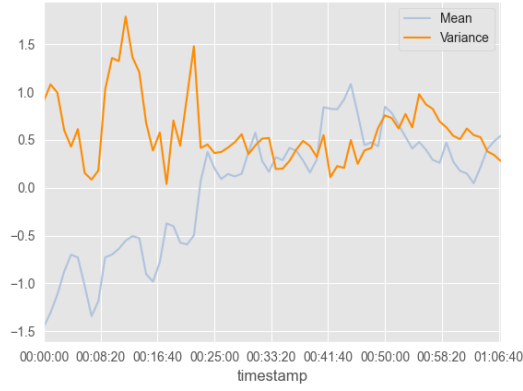


(o) Episode E19

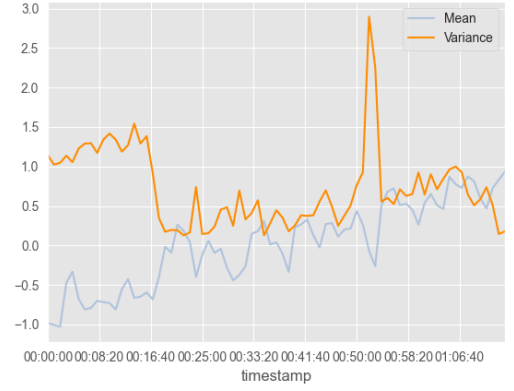


(p) Episode E15

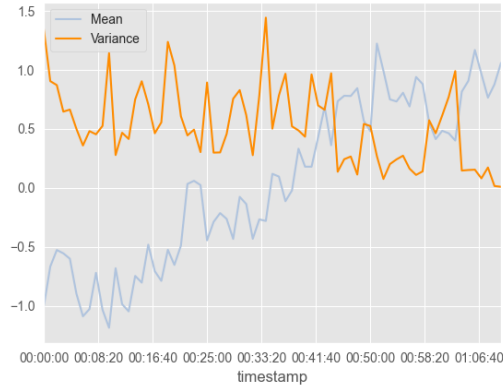
Figure 4.3: Mean and Standard Deviation between participants for airtime



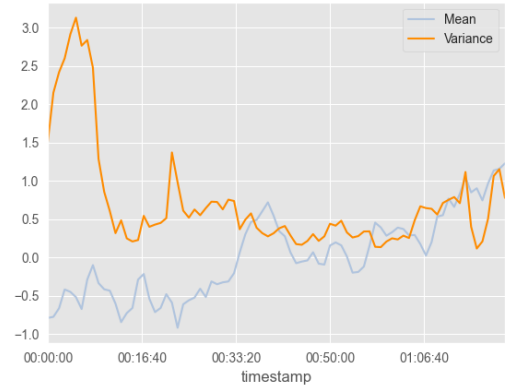
(q) Episode E16



(r) Episode E18

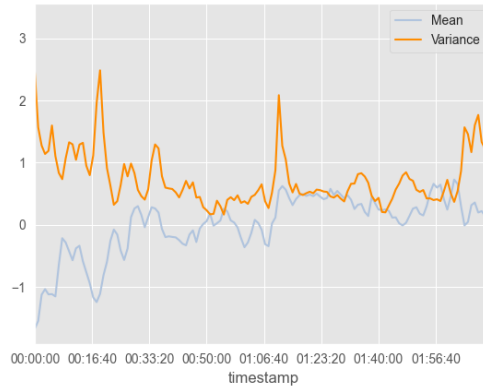


(s) Episode E20



(t) Episode E11

Figure 4.3: Mean and Standard Deviation between participants for airtime



(u) Episode E10

Figure 4.3: Mean and Standard Deviation between participants for airtime

D 5-Folder Cross Validation Results for XGBoost Model using EDA Components

Study	Mean MAE	Standard Deviation MAE
E21	0.950	0.225
E14	0.894	0.182
E19	0.863	0.139
E15	1.026	0.184
E16	0.880	0.107
E20	1.006	0.128
E11	0.975	0.192
E22	0.915	0.178
E10	0.866	0.132
E5	0.938	0.237
E38	0.791	0.095
E29	0.711	0.212
E24	0.700	0.118
E34	0.862	0.108
E35	0.868	0.137
E9	0.774	0.103
E39	0.662	0.169
E37	0.648	0.165
E36	0.923	0.160
E32	0.797	0.128
Average	0.852	0.155

Table 4.1: Results of the 5-fold cross-validation run: Average and standard deviation for XGBoost model using C2

Study	Mean MAE	Standard Deviation MAE
E21	1.284	0.412
E14	1.426	0.220
E19	1.343	0.235
E15	1.397	0.340
E16	1.253	0.224
E20	1.303	0.224
E11	1.366	0.264
E22	1.935	0.204
E10	2.268	0.386
E5	1.168	0.279
E38	2.212	0.166
E29	2.002	0.211
E24	2.497	0.278
E34	2.005	0.141
E35	1.887	0.225
E9	2.917	0.226
E39	2.663	0.172
E37	2.079	0.239
E36	1.498	0.348
E32	2.092	0.323
Average	1.829	0.255

Table 4.2: Results of the 5-fold cross-validation run: Average and standard deviation for MLP model using C2

Study	Mean MAE	Standard Deviation MAE
E21	0.915	0.196
E14	1.043	0.134
E19	0.973	0.103
E15	1.019	0.168
E16	0.884	0.178
E20	0.945	0.124
E11	1.007	0.163
E22	0.886	0.199
E10	0.967	0.112
E5	0.766	0.236
E38	0.850	0.127
E29	0.787	0.214
E24	0.767	0.115
E34	0.784	0.167
E35	0.795	0.143
E9	0.901	0.164
E39	0.657	0.141
E37	0.651	0.169
E36	0.886	0.115
E32	0.758	0.169
Average	0.862	0.157

Table 4.3: Results of the 5-fold cross-validation run: Average and standard deviation for XGBoost model using C3

Study	Mean MAE	Standard Deviation MAE
E21	1.101	0.127
E14	1.074	0.148
E19	1.043	0.219
E15	1.029	0.236
E16	0.964	0.185
E20	1.086	0.237
E11	1.127	0.231
E22	1.335	0.215
E10	1.396	0.188
E4	0.787	0.348
E38	1.477	0.249
E29	1.377	0.198
E24	1.480	0.279
E34	1.354	0.208
E35	1.378	0.307
E9	1.818	0.371
E39	1.747	0.344
E37	1.195	0.295
E36	1.005	0.193
E32	1.343	0.267
Average	1.263	0.242

Table 4.4: Results of the 5-fold cross-validation run: Average and standard deviation for MLP model using C3

Study	Mean MAE	Standard Deviation MAE
E21	0.918	0.197
E14	0.940	0.108
E19	1.121	0.197
E15	1.056	0.174
E16	0.859	0.199
E20	0.919	0.145
E11	0.896	0.198
E22	0.917	0.159
E10	0.877	0.133
E4	0.891	0.267
E38	0.825	0.164
E29	0.773	0.207
E24	0.735	0.081
E34	0.901	0.127
E35	0.882	0.146
E9	0.858	0.100
E39	0.693	0.149
E37	0.680	0.128
E36	0.889	0.125
E32	0.816	0.152
Average	0.872	0.158

Table 4.5: Results of the 5-fold cross-validation run: Average and standard deviation for XGBoost model using C4

Study	Mean MAE	Standard Deviation MAE
E21	1.060	0.115
E14	1.082	0.183
E19	1.058	0.205
E15	1.022	0.192
E16	0.944	0.187
E20	1.030	0.186
E11	1.110	0.205
E22	1.354	0.234
E10	1.395	0.202
E4	0.772	0.380
E38	1.455	0.219
E29	1.354	0.178
E24	1.495	0.263
E34	1.405	0.202
E35	1.378	0.274
E9	1.826	0.349
E39	1.745	0.346
E37	1.472	0.261
E36	1.072	0.203
E32	1.344	0.219
Average	1.269	0.230

Table 4.6: Results of the 5-fold cross-validation run: Average and standard deviation for MLP model using C4

E 5-Folder Cross Validation Results for XGBoost and MLP using the composite EDA Signal

Study	Mean MAE	Standard Deviation MAE
E21	1.453	0.341
E14	1.118	0.207
E19	1.352	0.218
E15	1.487	0.336
E16	1.352	0.285
E20	1.317	0.243
E11	1.265	0.229
E22	1.035	0.175
E10	1.071	0.134
E4	0.864	0.244
E38	0.902	0.112
E29	0.887	0.140
E24	0.804	0.089
E34	1.068	0.127
E35	1.035	0.181
E9	0.925	0.086
E39	0.788	0.163
E37	0.765	0.165
E36	1.344	0.273
E32	1.086	0.139
Average	1.096	0.194

Table 4.7: Results of the 5-fold cross-validation run: Average and standard deviation for XGBoost model using R2

Study	Mean MAE	Standard Deviation MAE
E21	1.067	0.186
E14	0.931	0.141
E19	1.022	0.194
E15	1.005	0.219
E16	1.054	0.257
E20	1.118	0.195
E11	1.027	0.170
E22	0.869	0.168
E10	0.814	0.085
E4	0.979	0.311
E38	0.767	0.124
E29	0.821	0.515
E24	0.718	0.092
E34	0.813	0.081
E35	0.807	0.133
E9	0.791	0.120
E39	0.679	0.177
E37	0.662	0.195
E36	1.021	0.205
E32	0.714	0.178
Average	0.714	0.178

Table 4.8: Results of the 5-fold cross-validation run: Average and standard deviation for MLP model using R2

Study	Mean MAE	Standard Deviation MAE
E21	1.167	0.175
E14	1.061	0.179
E19	1.035	0.179
E15	1.006	0.257
E16	0.924	0.164
E20	1.159	0.180
E11	1.047	0.211
E22	1.062	0.216
E10	1.261	0.175
E4	0.815	0.248
E38	0.993	0.164
E29	1.039	0.201
E24	0.904	0.127
E34	0.941	0.192
E35	0.956	0.153
E9	0.965	0.174
E39	0.740	0.139
E37	0.804	0.185
E36	1.000	0.293
E32	0.875	0.174
Average	0.988	0.174

Table 4.9: Results of the 5-fold cross-validation run: Average and standard deviation for XGBoost model using R3

Study	Mean MAE	Standard Deviation MAE
E21	0.843	0.143
E14	0.781	0.165
E19	0.865	0.127
E15	0.874	0.166
E16	0.763	0.163
E20	0.926	0.149
E11	0.819	0.178
E22	0.736	0.167
E10	0.751	0.109
E4	0.583	0.280
E38	0.665	0.113
E29	0.631	0.172
E24	0.571	0.119
E34	0.737	0.127
E35	0.733	0.129
E9	0.687	0.130
E39	0.556	0.158
E37	0.525	0.166
E36	0.785	0.170
E32	0.622	0.147
Average	0.722	0.147

Table 4.10: Results of the 5-fold cross-validation run: Average and standard deviation for MLP model using R3

Study	Mean MAE	Standard Deviation MAE
E21	1.181	0.255
E14	1.069	0.171
E19	1.067	0.177
E15	1.179	0.209
E16	0.917	0.184
E20	1.051	0.203
E11	1.064	0.172
E22	1.096	0.206
E10	1.202	0.167
E4	0.698	0.323
E38	0.973	0.204
E29	1.034	0.140
E24	0.925	0.148
E34	1.036	0.249
E35	0.987	0.164
E9	1.066	0.140
E39	0.885	0.197
E37	0.740	0.133
E36	0.899	0.214
E32	0.878	0.134
Average	0.998	0.0.19

Table 4.11: Results of the 5-fold cross-validation run: Average and standard deviation for XGBoost model using R4

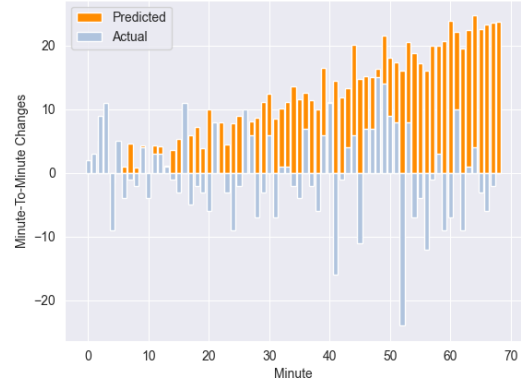
Study	Mean MAE	Standard Deviation MAE
E21	0.777	0.140
E14	0.780	0.120
E19	0.865	0.169
E15	0.914	0.146
E16	0.766	0.188
E20	0.847	0.098
E11	0.821	0.174
E22	0.730	0.168
E10	0.751	0.106
E4	0.602	0.304
E38	0.657	0.107
E29	0.646	0.218
E24	0.569	0.117
E34	0.737	0.133
E35	0.717	0.142
E9	0.668	0.132
E39	0.552	0.158
E37	0.507	0.157
E36	0.721	0.167
E32	0.620	0.171
Average	0.712	0.155

Table 4.12: Results of the 5-fold cross-validation run: Average and standard deviation for MLP model using R4

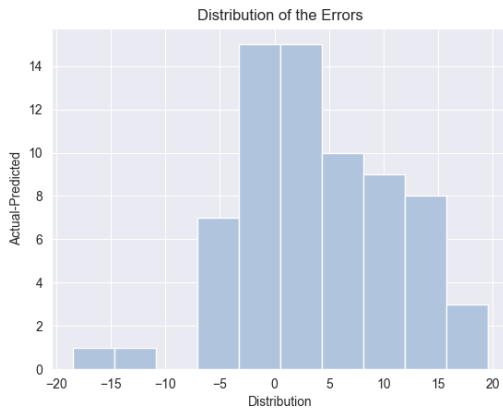
F Visualisation of the errors for MLP model



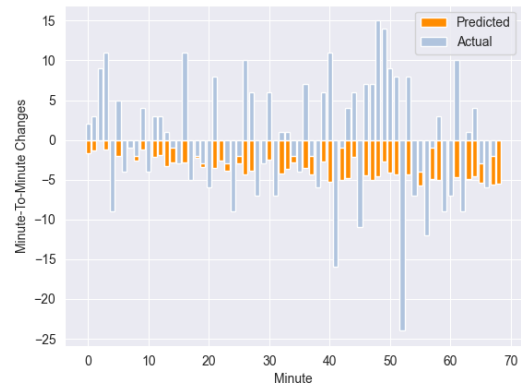
(a) Histograms of Error for MLP on R1



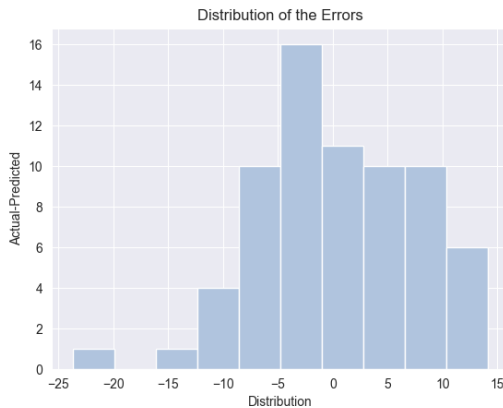
(b) Actual and Predicted values for MLP on R1



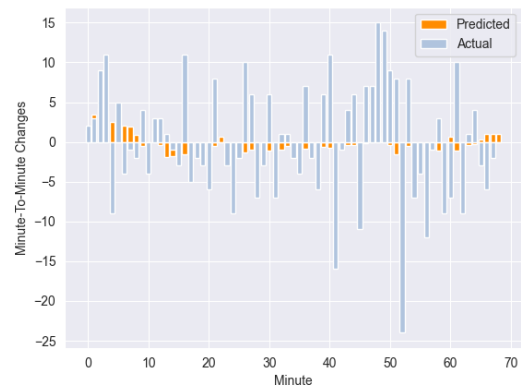
(c) Histograms of Error for MLP on R2



(d) Actual and Predicted values for MLP on R2

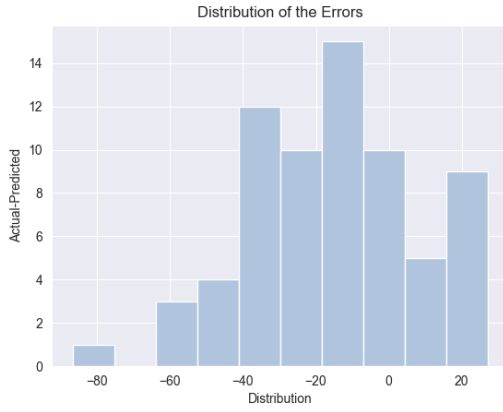


(e) Histograms of Error for MLP on R3

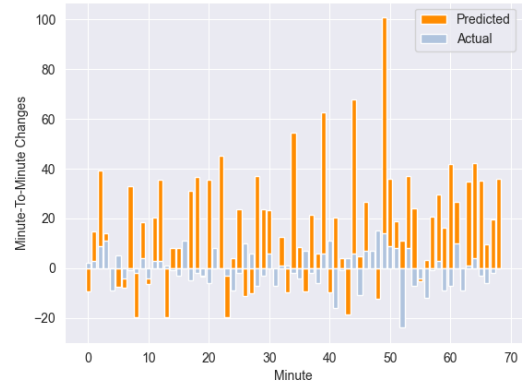


(f) Actual and Predicted values for MLP on R3

Figure 4.4: Prediction of Twitter Activity for MLP Model



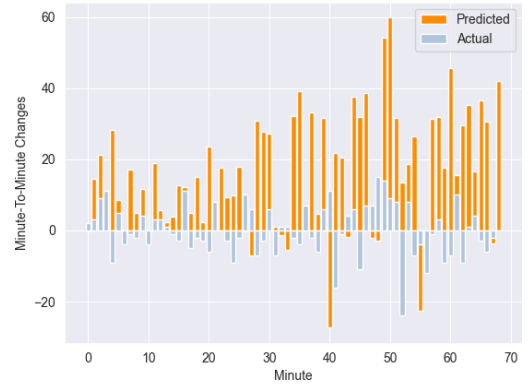
(g) Histograms of Error for MLP on C1



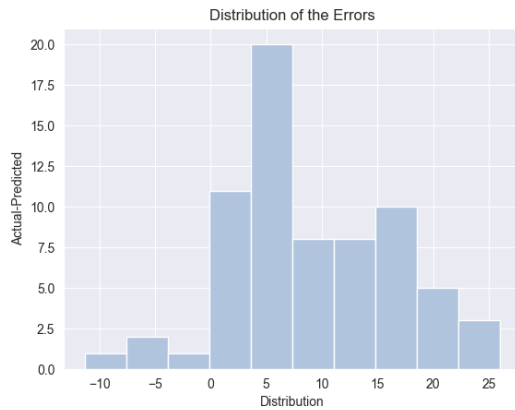
(h) Actual and Predicted values for MLP on C1



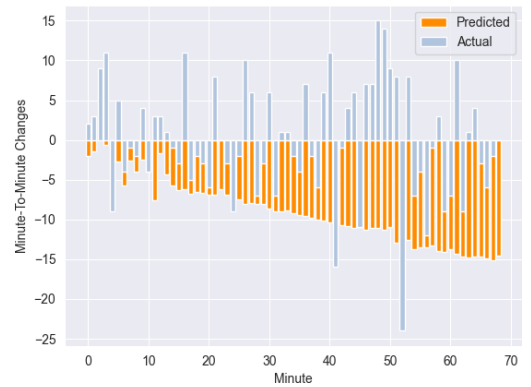
(i) Histograms of Error for MLP on C2



(j) Actual and Predicted values for MLP on C2

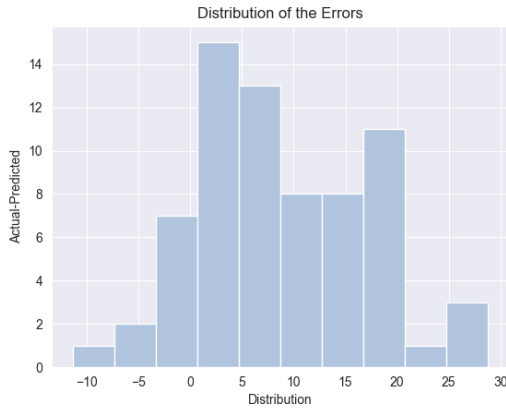


(k) Histograms of Error for MLP on C3

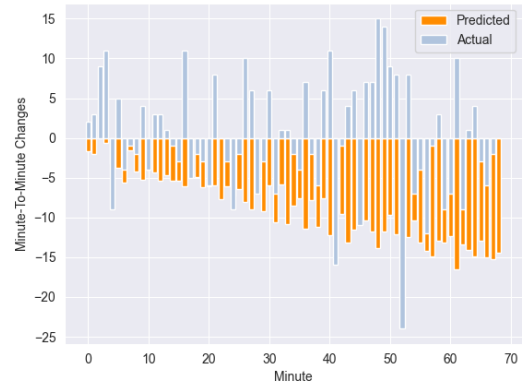


(l) Actual and Predicted values for MLP on C3

Figure 4.4: Prediction of Twitter Activity for MLP Model



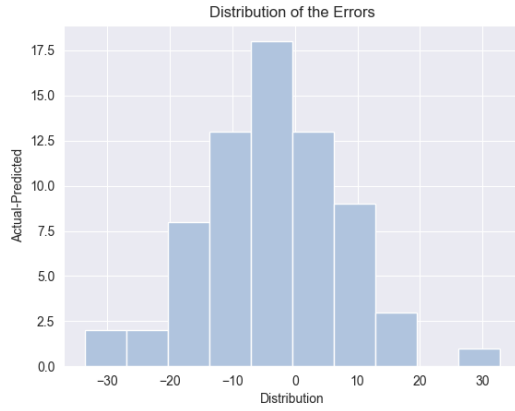
(m) Histograms of Error for MLP on C4



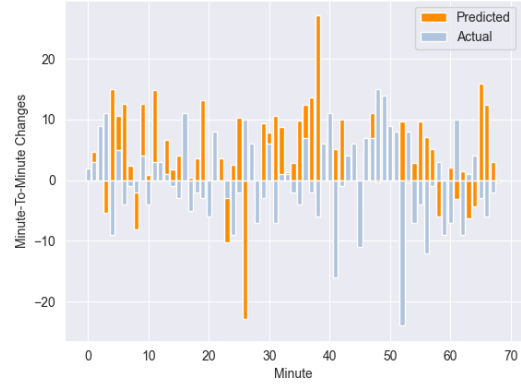
(n) Actual and Predicted values for MLP on C2

Figure 4.4: Prediction of Twitter Activity for MLP Model

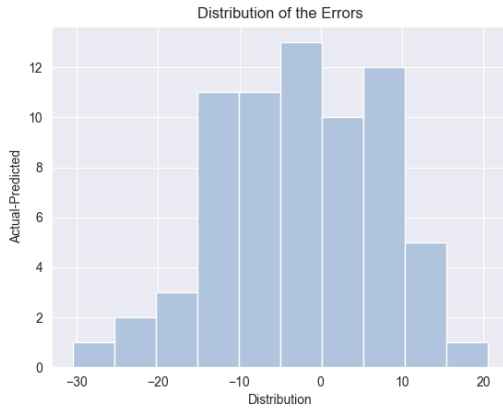
G Visualisation of the errors for XGBoost model



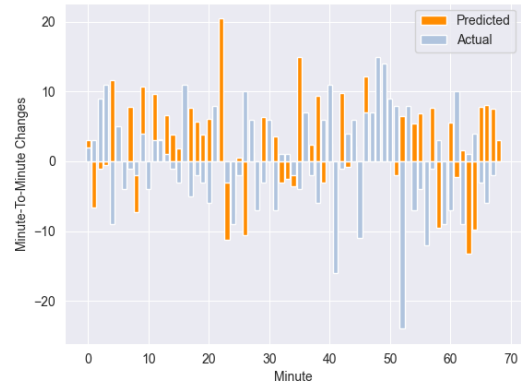
(a) Histograms of Error for XGBoost on R1



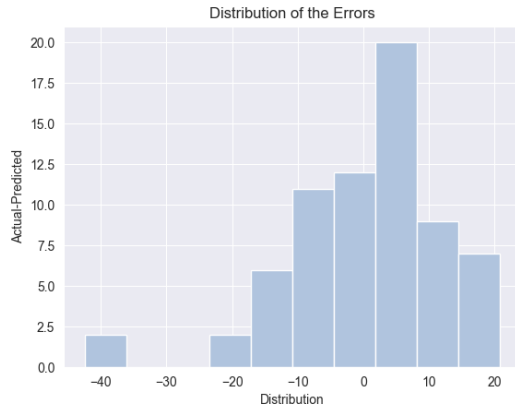
(b) Actual and Predicted values for XGBoost on R1



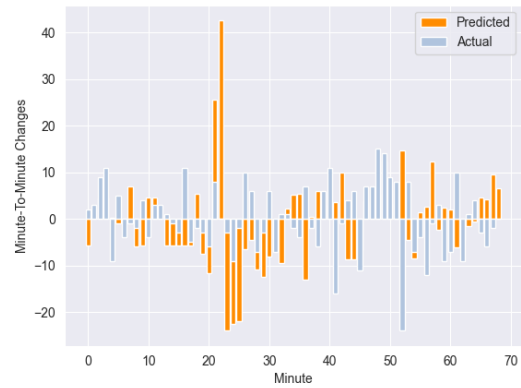
(c) Histograms of Error for XGBoost on R2



(d) Actual and Predicted values for XGBoost on R2

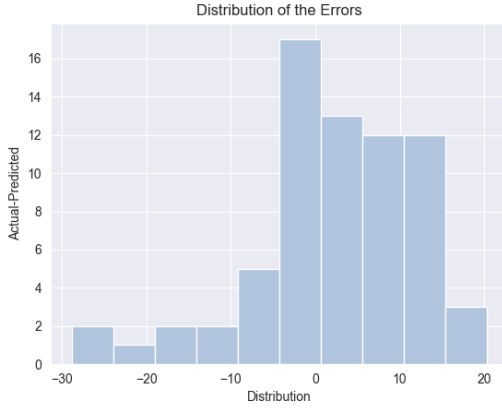


(e) Histograms of Error for XGBoost on R3

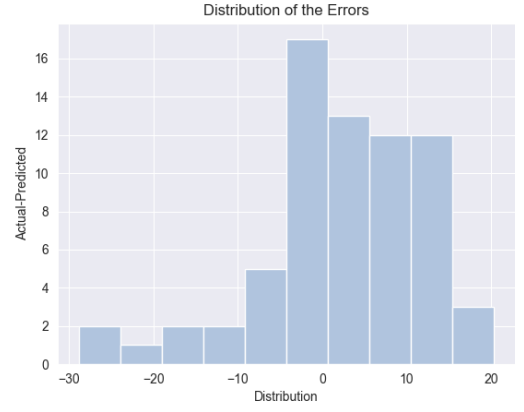


(f) Actual and Predicted values for XGBoost on R3

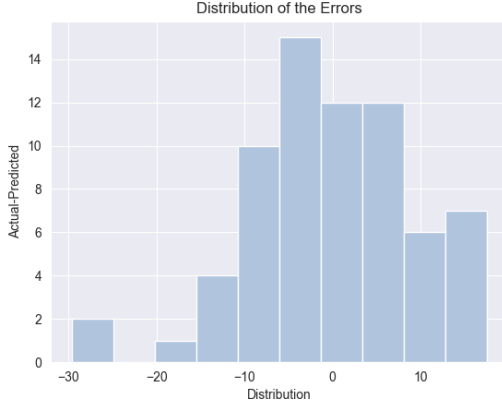
Figure 4.5: Prediction of Twitter Activity for XGBoost Model



(g) Histograms of Error for XGBoost on R4



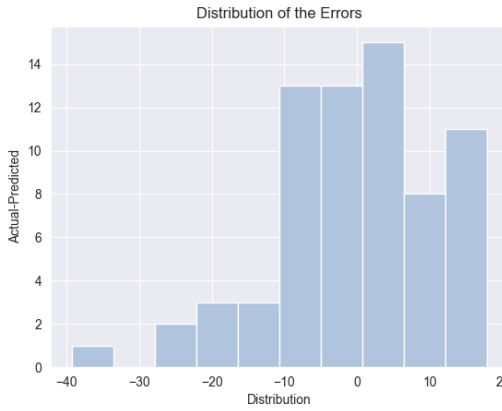
(h) Actual and Predicted values for XGBoost on R4



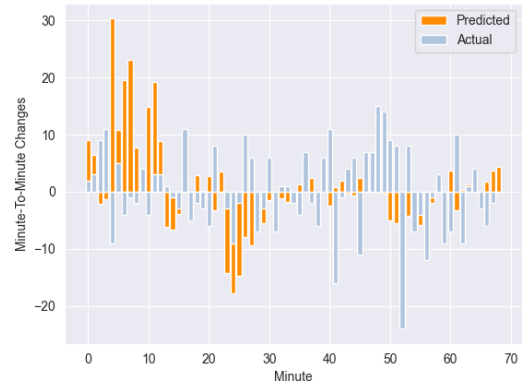
(i) Histograms of Error for XGBoost on C1



(j) Actual and Predicted values for XGBoost on C1

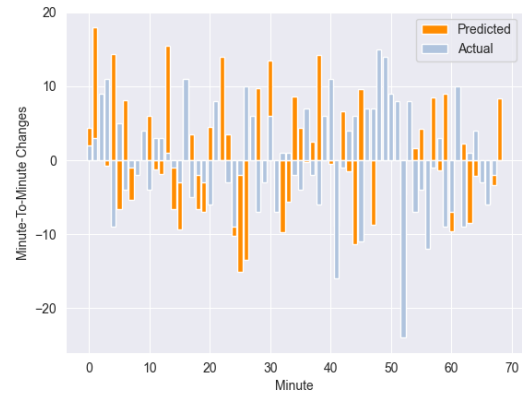
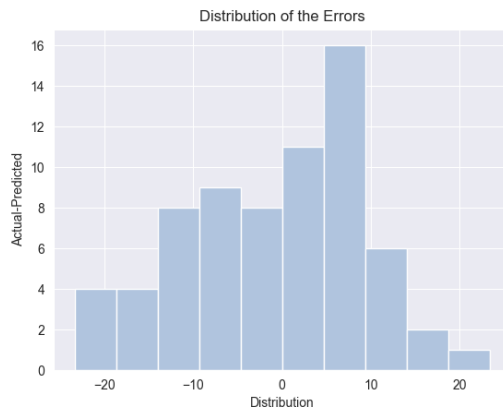


(k) Histograms of Error for XGBoost on C3



(l) Actual and Predicted values for XGBoost on C3

Figure 4.5: Prediction of Twitter Activity for XGBoost Model



(m) Histograms of Error for XGBoost on C4

(n) Actual and Predicted values for XGBoost on C2

Figure 4.5: Prediction of Twitter Activity for XGBoost Model