UNIVERSITA' DEGLI STUDI DI PADOVA

DIPARTIMENTO DI SCIENZE ECONOMICHE ED AZIENDALI "M.FANNO"

CORSO DI LAUREA MAGISTRALE IN ECONOMICS AND FINANCE

TESI DI LAUREA

"The Pandemic Effect on the Public Perception of Science: Evidence from Documentary Reviews"

RELATORE:

CH.MO PROF. LUCA NUNZIATA

LAUREANDO: MATEUS GOMES DA SILVA RODRIGUES

MATRICOLA N. 2051147

ANNO ACCADEMICO 2022 – 2023

Il candidato dichiara che il presente lavoro è originale e non è già stato sottoposto, in tutto o in parte, per il conseguimento di un titolo accademico in altre Università italiane o straniere.
Il candidato dichiara altresì che tutti i materiali utilizzati durante la preparazione dell’elaborato sono stati indicati nel testo e nella sezione “Riferimenti bibliografici” e che le eventuali citazioni testuali sono individuabili attraverso l’esplicito richiamo alla pubblicazione originale.

The candidate declares that the present work is original and has not already been submitted, totally or in part, for the purposes of attaining an academic degree in other Italian or foreign universities. The candidate also declares that all the materials used during the preparation of the thesis have been explicitly indicated in the text and in the section "Bibliographical references" and that any textual citations can be identified through an explicit reference to the original publication.

Firma dello studente

Mateus Rodrigues

Abstract

This study aims to estimate the effect of the COVID-19 Pandemic on anti-scientific sentiments, exploiting movie reviews as a metric of public opinion. Using a Differences-in-Differences approach, we test whether the Pandemic led to a change in the general public's perception of scientific documentaries. We find that the pandemic led to discernible shifts in the public's receptivity to scientific content. Indeed, ratings for scientific documentaries decreased by an additional 0.47-0.76 following the onset of the public health emergency, relative to other movies. Similarly, the pandemic led to a further 0.028-0.033 increase in the probability of a science-related documentary review having an angry sentiment, when compared to other genres. These shifts suggest that the pandemic may have contributed to the emergence or expansion of an increasingly popular anti-science movement, characterized by the disbelief of science and the scientific process.

Keywords: anti-science, COVID-19, pandemic, emotion detection, movies reviews

Contents

Chapter 1

Introduction

In the wake of the COVID-19 pandemic, a remarkable phenomenon emerged: the rise of movements characterized by anti-scientific sentiments and the propagation of scientific misinformation. In this study, we aim to identify and analyze whether these behaviors and propagandistic tendencies during the pandemic can be identified and quantified through the examination of scientific documentary reviews. Our research seeks to shed light on the extent to which public perceptions of science may have been influenced by such movements during this unique period of global upheaval.

This paper attempts to unravel the connections between the pandemic and public perceptions of science by examining documentary viewers' reactions, their evolving evaluations of scientific content, and the emotions conveyed in their reviews. We will examine whether the pandemic acted as a catalyst for increased interest in science, or whether it led to new forms of skepticism and criticism.

To address these questions, we use a differences-in-differences model. In our study, we focus primarily on evaluating two different aspects of the reviews: their numerical ratings and the presence of expressions of anger within them. The rating component reflects users' ratings on a scale from 1 to 10. To analyze the emotional state, specifically the presence of anger, we will use a machine learning algorithm in this paper. The process of this emotion classification task will be explored in more detail in the following chapters.

The reviews we use for this analysis were created exclusively for this paper. For both the movies and the science documentaries, we use data from two of the largest English-language review sites: IMDB and Rotten Tomatoes. In the following chapters, we will go into further details on the specific criteria used to construct this new dataset.

With this analysis, we aim to offer insights into the transformation of science in the public eye and the potential impacts that a shock like the Pandemic can have on individuals' response to the scientific knowledge.

Chapter 2

Context

2.1 Theoretical Framework

In order to analyze the effect of the COVID-19 pandemic on attitudes towards science, it is first necessary to think about the possible mechanisms through which this Public health emergency might affect public perceptions. First, by directly increasing the attention given to scientific matters. Second, by serving as a catalyzer for the pre-existing populist trend, which in turn can affect societal beliefs about science.

The mechanism through which populists movements could impact mass opinions on science has been an extensive topic of interest. In Mede et al. (2021) the authors discuss the relation of populist movements and science. They define populism as a political ideology that pits "the people" against "the elite." They argue that science-related populism is a new, specific type of populism that targets scientists and the scientific establishment. Science-related populists often view scientists as out-of-touch with the concerns of ordinary people and as being motivated by self-interest rather than the public good. They may also distrust scientific expertise and promote alternative sources of knowledge, such as personal experience or common sense.

Another important change, which began to occur in the 1960s, was identified by social scientists is the decline in traditional forms of citizen or lay participation, such as elections (Kostelka, 2017), giving space to nontraditional forms of participation in many realms of society, a process which was labeled “participatory turn”. This phenomenon” refers to, among other things, the increasing emphasis on public participation in scientific decision-making. The “participatory turn” opened the space traditionally destined to the scientific debate to alternative epistemologies, such as traditional knowledge.

Mede et al. (2021) argue that these trends have created a more open and contested space for science, which has made it more vulnerable to populist attacks.The authors then propose a definition of science-related populism as a set of ideas that frames the relationship between scientists and the public as an antagonistic one. They argue that science related populism is characterized by some core beliefs, like:

1. Scientists are out of touch with the concerns of ordinary people.
2. Scientists are motivated by self-interest rather than the public good.
3. Scientific expertise is unreliable and should be distrusted.
4. Alternative sources of knowledge, such as personal experience or common sense, are equally valid or even more valid than scientific knowledge.

2.2 Existing Evidence

Evidence suggests that adverse macroeconomic shocks can create a demand for populism in Europe. Gavresi et al. (2023) find that when individuals are faced with recessions during their earlier adult life, they tend to participate less in the political scene of their country, and the ones who do, tend to support populist parties or leaders. These individuals also tend to have a lower trust in the government.

Empirical evidence of a relationship between anti-science sentiments and populist movements, as predicted by Mede et al. (2021), also exist. Rao et al. (2021) examine the relationship between political partisanship and these anti-science attitudes in online discussions about COVID-19, using a dataset of over 4 million tweets collected during the early months of the COVID-19 pandemic in the United States.

They find that political partisanship is strongly associated with anti-science attitudes in online discussions about COVID-19. Republicans are more likely than Democrats to express anti science views, even after controlling for other factors such as age, gender, education, and income. Anti-science attitudes are more likely to be expressed in certain states. The authors found that anti-science attitudes are more common in states that voted for Donald Trump in the 2016 presidential election. They also found that anti-science attitudes are associated with a variety of other outcomes. People who express anti-science views are more likely to have lower levels of trust in science and public health institutions, and they are more likely to be vaccine hesitant.

The authors argue that these findings suggest that political partisanship is a major factor driving the spread of misinformation and disinformation about COVID-19 and science in general. This research sheds more light on the direct and substantial relationship between the populist movement, strongly influenced by Trump in the United States, and the prevalence of anti-scientific beliefs among its followers.

In summary, evidence exists that adverse shocks can potentially lead to polarization and to the rise of populist movements (Gavresi et al, 2023). Furthermore, Rao et al. (2021) offer insights into how the polarization and political identification with populist parties is correlated to anti-scientific views. Although it is expected that an adverse shock like the COVID-19 pandemic could potentially lead to increased resentment and anger directed at the scientific community and scientific principles, little is known about the true effects of this event on political views and sentiment towards science.

The contributions of this paper are twofold. First, it provides concrete information on how the pandemic affected public opinions on scientific content. Second, it shows that movie reviews can accurately be used to provide insight on public sentiments on politically loaded questions. In the context of a trend towards decreased public availability of data sources reflecting real-time user perceptions, with Twitter and Reddit APIs closing down, this is a particularly important contribution for research in social sciences.

Chapter 3

Natural Language Processing

Natural Language Processing (NLP) is the realm of developing machinery capable of comprehending and manipulating human language or language-like data, encompassing written, spoken, and organized forms. This discipline originated from computational linguistics, a field employing computer science to unravel linguistic principles. However, NLP distinguishes itself by focusing on the practical engineering aspects of creating technology to perform useful tasks.

NLP comprises two interrelated subfields: Natural Language Understanding (NLU), which delves into semantic analysis and discerning the intended meaning within text, and Natural Language Generation (NLG), dedicated to machine-based text creation. While NLP is separate from speech recognition, it is often intertwined with it to transcribe spoken language into text and vice versa.

NLP plays a pivotal role in our daily lives and continues to gain importance as language technology permeates various sectors, including retail and healthcare. Conversational agents like Amazon's Alexa and Apple's Siri utilize NLP to comprehend user queries and provide responses. The most advanced agents, such as GPT-3, are capable of generating sophisticated prose on diverse subjects and empowering chatbots to hold coherent conversations.

Leading technology companies like Google employ NLP to enhance search engine results, while social networks like Facebook use it to identify and filter hate speech.

Natural Language Processing has garnered immense attention due to its versatile applications, including text generators capable of crafting coherent essays, chatbots that mimic sentience, and text-to-image programs that can generate photorealistic representations of described concepts.

Recent years have seen computers undergo a transformative change, capable of understanding not only human languages but also programming languages, complex biological and chemical sequences resembling linguistic structures. AI models' latest advancements have unlocked its potential to explore these areas; deciphering input text's meanings and producing meaningful and expressive output.

3.1 Emotions

Emotion recognition is a language-related task inside the realms of NPL. When dealing with emotion recognition, there are usually two common approaches to determine the set of emotions that will be used, Plutchik's wheel of emotions or Ekman's six basic emotions.

Plutchik's wheel of emotions is a model of human emotions that organizes eight primary emotions and their related emotions into a circular pattern. The wheel of emotions is divided into four pairs of opposite emotions, with each pair representing a different dimension of emotion. For example, joy and sadness are opposites on the dimension of valence (positive vs. negative), while anger and fear are opposites on the dimension of intensity (high vs. low).

Plutchik also believed that the primary emotions can be combined to form more complex emotions. For example, the combination of joy and trust results in love, while the combination of anger and disgust results in hatred. Plutchik argues that these emotions are universal across cultures and innate and have a set of characteristics like being identifiable through facial expressions and other nonverbal cues, having a distinct physiological profile and that they all serve important adaptive functions.

In this paper, the Plutchik’s Wheel of Emotions will be utilized.

3.2 Detecting Emotions in Texts

3.2.1 Lexicon Based Approach

A lexicon-based approach can be utilized to determine the emotional content of a text. This technique is based on the presence and frequency of specific words from predefined emotion lexicons. This approach relies on the assumption that certain words are strongly associated with particular emotions. Lexicons are dictionaries or lists of words that have been manually or automatically labeled with specific emotions, each word in the lexicon is associated with one or more emotions.

To perform emotion recognition using lexicons, the process involves analyzing the input text, such as a movie review, to detect the presence of emotion words from the lexicon. The frequency and context of these words can also be considered. Once the emotion-associated words are identified in the text, they can be assigned scores or weights based on their relevance or intensity. For example, a word like "ecstatic" may contribute more to the ’joy’ score than a less intense word like "happy."

In Strapparava et al (2008), the authors implemented an algorithm that would check lexical and semantic features in news headlines. It achieved good results in fine-grained emotion identification. Their method was able to identify the correct emotion category for a headline with an accuracy of around 80%.

In Balahur et al (2011) EmotiNet was introduced, a knowledge base of concepts with associated affective values. EmotiNet is a knowledge base that contains over 10,000 concepts with associated affective values. Each concept in EmotiNet is annotated with a set of affective labels, such as happiness, sadness, anger, fear, surprise, and disgust.

The authors then used EmotiNet to develop a method for detecting implicit expressions of sentiment in text. Balahur et al. (2011) argue that commonsense knowledge can be used to detect implicit expressions of sentiment. The authors evaluated their method on a dataset of news articles and found that it was able to detect implicit expressions of sentiment with an accuracy of over 70%.

Another common iteration of the lexicon approach is the use of Latent Semantic Analysis (LSA). LSA is a technique of analyzing relationships between a set of corpus and the terms contained within it, by producing a set of concepts related to the corpus and terms, creating a set of patterns. In Gill et al (2008) a combination of LSA and Hyperspace Analogue to Language (HAL) was used to calculate the semantic similarity between texts and emotions keywords from blogs texts.

3.2.2 Machine Learning Based Approach

The Machine learning approach, which was used in this paper, involves training models to automatically identify and classify emotions within textual data. These methods do not rely on predefined word lists, but rather learn patterns and relationships to detect emotion.

Specifically in supervised machine learning approaches, the algorithms depend on a labeled training data, which then the model can infer a function, which can be used for mapping new unlabeled data. The labeling processing is usually manually annotated by humans and, although a very time-consuming task, represents an essential step into deploying a successful machine learning model. In the last few years, there have been works in the sphere of NLP that use automatics labeling via a collection of hashtags in Twitter messages. In Saravia et al (2018) a set of hashtags was constructed to collect a dataset of English tweets from the Twitter API, considering the hashtag appearing in the last position of a tweet as the ground truth.

When dealing with supervised learning algorithms, it is usual to find both the categorical and the dimensional approaches. Categorical approach is the most commonly used in emotion recognition, it classifies text into predefined sentiment categories. This approach simplifies sentiment assessment by reducing the complexity of emotions and opinions expressed in text to a few discrete labels. It is often employed in applications where a quick determination of sentiment is sufficient.

Conversely, the dimensional approach can be seen in work like Hasan et al (2014), in which they propose an automatic classifier for text messages to identify their emotional states. They use the Rusell's Circumflex Model of Affect as an emotion model and train the model to detect multiple emotions within the text.

Chapter 4

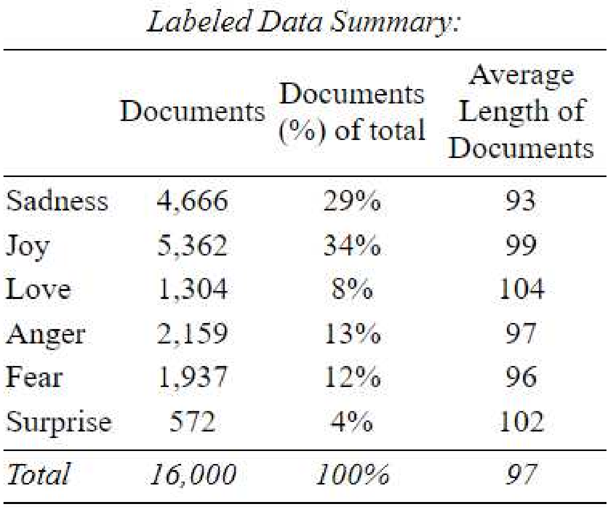
Dataset

4.1 Dataset Description

4.1.1 Labeled Data

For the pre-labeled dataset, we use the data gathered by Saravia et al (2018). The authors constructed a set of hashtags to collect a dataset of English tweets from the Twitter API belonging to six emotions, including sadness, joy, love, anger, fear and surprise.

The authors then manually reviewed 16,000 unique tweets and assigned the appropriate emotion. Some basic numbers from this dataset are shown below:



Labeled Data Summary:

	Documents	Documents (%) of total	Average Length of Documents
Sadness	4,666	29%	93
Joy	5,362	34%	99
Love	1,304	8%	104
Anger	2,159	13%	97
Fear	1,937	12%	96
Surprise	572	4%	102
Total	*16,000*	*100%*	*97*

Figure 4.1: Labeled Data Summary

The pre-labeled dataset used in this paper has already been preprocessed based on the approach described in their paper.

4.1.2 Movies Reviews Data

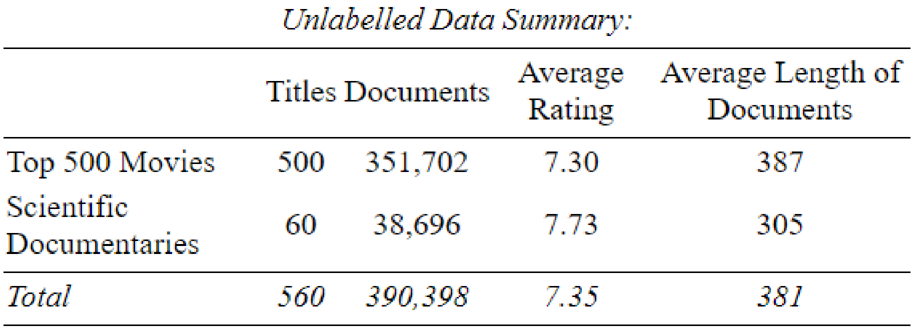
This paper collected data from two prominent English-language movie review platforms, namely the Internet Movie Database (IMDB) and Rotten Tomatoes, using a custom web scraping algorithm to create two distinct datasets.

The first dataset was constructed by selecting the top 500 titles with the highest number of user reviews on IMDB, all of which were released between 2010 and 2019. The selection of 500 movies was made with the belief that this quantity accurately represents what is popular worldwide. This notion is reinforced by the dataset's inclusion of a diverse array of movie genres, further solidifying its representativeness. You can find a graph depicting the distribution of reviews by movie genre in Annex I. The dataset encompasses 19 distinct genres and exhibits a well-balanced distribution among them.

User-written reviews for the 500 movies were scraped from both IMDB and Rotten Tomatoes, resulting in a dataset of 351,702 reviews.

For the second dataset, our focus was on documentaries with a scientific emphasis, covering topics such as climate change, environmental issues, biology, and medicine. We utilized IMDB's keyword search functionality, which allows users to tag movies with specific keywords that they believe apply to that movie. These keywords are then rated by other users on the basis of their "usefulness". By searching for keywords like "science," "nature," and "environment," we compiled a list of pertinent documentaries released between 2010 and 2019. This dataset consists of 60 documentaries. The number 60 was selected due to a lack of written user reviews for periods preceding and following the pandemic beyond number 61. To further expand our review pool, we cross-referenced this list of documentaries in Rotten Tomatoes.

Some basic numbers from these datasets are shown below:



Unlabelled Data Summary:

	Titles	Documents	Average Rating	Average Length of Documents
Top 500 Movies	500	351,702	7.30	387
Scientific Documentaries	60	38,696	7.73	305
Total	*560*	*390,398*	*7.35*	*381*

Figure 4.2: Unlabeled Data Summary

Our novelty dataset comprises a total of 390,398 reviews, encompassing 560 distinct movie titles. Among these, 500 reviews stem from what can be argued as the most prominent movies from the period spanning 2010 to 2019 in terms of popularity. Additionally, there are 60 science documentaries included in the dataset.

Interestingly, the scientific documentaries exhibit a higher average rating when compared to the top 500 movies. Furthermore, on average, the reviews associated with scientific documentaries tend to contain fewer characters.

4.2 Data Pre-Processing

Data pre-processing is a critical and foundational step in the field of machine learning. It involves a series of operations and techniques applied to raw data before it can be effectively used to train machine learning models. The primary objectives of data pre-processing are to improve data quality, ensure data compatibility with the chosen algorithms, and enhance the overall model's performance.

4.3 Tokenization

Tokenization is the process of dividing a text into smaller, meaningful units called tokens. The result generally consists of tokenized text in which words may be represented as numerical tokens for various uses.

These tokens are the building blocks in the NLP, most of the preprocessing and modeling happens at a token level.

4.3.1 Feature Extraction

Once tokenization is complete, the next step in emotion recognition involves feature extraction. This process transforms the tokens into numerical representations that can be used for machine learning algorithms. Two common feature extraction techniques used in emotion recognition are Bag-of-Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF).

Bag-Of-Words (BoW)

BoW is a simple and effective method for feature extraction. It represents each document as a vector of word frequencies. The idea is to create a vocabulary of all unique words in the entire dataset and count how often each word appears in a document.

In the context of emotion recognition, BoW can be used to create a document-term matrix where each row represents a document and each column represents a unique word. The cell values indicate the frequency of each word's occurrence in the document. While BoW discards word order and grammar, it provides a straightforward way to capture the emotional tone conveyed by specific words.

Term Frequency - Inverse Document Frequency

TF-IDF is a more advanced feature extraction method that combines term frequency (TF) and inverse document frequency (IDF).

In the context of natural language processing, the TF-IDF vectorization process is a pivotal step in converting textual data into a numerical representation. This technique facilitates the transformation of a collection of text documents into a structured matrix. During this process, the vocabulary of terms is established, and each document's term frequency is evaluated. Importantly, TF-IDF takes into account not only how frequently a term occurs within a specific document, but also how unique it is across the entire corpus of documents.

This process yields a high-dimensional matrix where each document is represented as a vector, with each element of the vector corresponding to the TF-IDF weight of a specific term within that document. The result is a numerical representation that encapsulates the semantic content of the text data, enabling further analysis and machine learning applications without the need for specific implementation details or library references.

Mathematically we have:

$$TF_{ij} = \frac{f_{ij}}{n_j}$$

Where fij is the frequency of a term i in document j, n_j is the total number of words in document j.

$$IDF_i = 1 + \log\left(\frac{N}{c_i}\right)$$

Where N is the total number of documents in the corpus. c_i is the number of documents that contain word i.

$$w_{ij} = TF_{ij} * IDF_i$$

Where the term w_{ij} is the TF-IDF score of term i in document j.

When computing the TF-IDF for every word in a document, we can generate a matrix with the shape ’number of words’ x ’number of documents’ The TF-IDF gives a single value for one word, but a matrix of values when considering the many documents.

Chapter 5

Classifiers

In this paper, we employed four distinct classifiers: Logistic Regression, Support Vector Machines, Random Forest, and Naive Bayes. These choices were made based on their well-established effectiveness in numerous text classification studies and their proven track record in a wide range of applications.

5.1 Logistic Regression

The term "logistic regression" itself was coined by Joseph Berkson in 1944 when he described a method for analyzing binary data. But the method gained more prominence in the late 20th century, notably through the work of David A. Cox and Emanuel Parzen. The development of computing technology and its growing accessibility in the mid-20th century facilitated the practical implementation of logistic regression for classification tasks.

Logistic regression's significance in the realm of machine learning and data science was further solidified through the groundbreaking work of Leo Breiman and Jerome Friedman in the 1980s. Their research on classification and regression trees, which are closely related to logistic regression, brought attention to the broader field of classification algorithms and paved the way for its application in various domains, including economics, medicine, and social sciences.

Logistic regression is used primarily for binary classification tasks, although it can be extended to handle multi-class classification as well. Unlike linear regression, which predicts continuous values, logistic regression models the probability of an input belonging to one of the classes. It employs the logistic function, also known as the sigmoid function, to transform a linear combination of input features into a value between 0 and 1, which can be interpreted as a probability.

Logistic regression can be applied to emotion recognition tasks by mapping input features, like text sentiment, to the probability of a particular emotion being present. In the context of emotion recognition, it serves as a simple yet effective model for classifying emotions into categories. By training on labeled datasets that associate input data with specific emotions, logistic regression learns the relationships between input features and the likelihood of each emotion. This allows the model to make probabilistic predictions, indicating the probability of an individual expressing a given emotion based on their input data.

The multi-class logistic regression model can be expressed as follows for each class k:

$$P(Y = k) = \frac{e^{\beta_{0k}+\beta_{1k}X_1+\beta_{2k}X_2+\ldots+\beta_{nk}X_n}}{\sum_{j=1}^{K} e^{\beta_{0j}+\beta_{1j}X_1+\beta_{2j}X_2+\ldots+\beta_{nj}X_n}}$$

Where K is the number of classes (in our case, $K = 6$).

$P(Y = k)$ is the probability that the text belongs to class K.

X is a vector of text features, we can use different techniques for creating this vector, such as TF-IDF or word embeddings.

β_{0k} is the intercept for class K and $\beta_{1k}, \beta_{2k}, \ldots, \beta_{nk}$ are the coefficients associated with each feature in X for class K.

The coefficients $\beta_{0k}, \beta_{1k}, \beta_{2k}, \ldots, \beta_{nk}$ are estimated from the training data. The class with the highest probability after applying the function is the predicted emotion class for the text.

5.2 Support Vector Machine

The concept of SVM dates back to the early 1960s when Vladimir Vapnik, a Russian mathematician and computer scientist, began exploring the idea of a "maximally flat" decision boundary for binary classification problems. This work laid the theoretical foundation for what would later become the SVM. Vapnik, along with Alexey Chervonenkis, introduced the Vapnik-Chervonenkis (VC) dimension, a critical concept in the analysis of the model's capacity to generalize from training data.

The modern formulation of SVM for classification tasks, known as the "hard-margin" SVM, was introduced by Vapnik et al. (1995). This paper presented the idea of finding a hyperplane that maximizes the margin between two classes in a binary classification problem. The margin is defined as the distance between the hyperplane and the nearest data points of each class. This concept led to the development of the mathematical optimization problem that SVM seeks to solve.

Support Vector Machine (SVM) is a powerful and widely used machine learning algorithm for classification and regression tasks. It operates by finding the optimal hyperplane that best separates data points in a high-dimensional feature space. The key idea behind SVM is to identify a decision boundary that maximizes the margin between different classes, where the margin is the distance between the hyperplane and the nearest data points, known as support vectors. SVM is effective in handling both linearly separable and non-linearly separable data, thanks to kernel functions that can transform the feature space to a higher dimension.

The formula for SVM in binary classification is as follows:

$$h_\theta(x) = \theta^T x + b$$

Here, $h_\theta(x)$ is the decision boundary $theta$ is the weight vector, x is the feature vector for the text, and b is the bias term.

For multi-class classification, it is practice to use the one-vs-all approach. You train K binary classifiers, one for each emotion. For the k-th classifier (where $k = 1, 2, \ldots, K$) we transform the problem into a binary classification problem where class k is treated as the positive class, and all other classes are treated as the negative class.

To predict the class for a given text, you apply all K binary classifiers to the text and choose the class that maximizes $h^{(k)}(x)$:

$$y = \arg\max_{k} h^{(k)}(x)$$

The choice of the kernel function also plays a major role, as it determines how data is mapped into a higher-dimensional. The linear kernel is the simplest of all SVM kernels. It computes the dot product between two feature vectors in the original feature space. The polynomial kernel transforms the data into a higher-dimensional space using polynomial functions of the original features. The Radial Basis Function kernel, also known as the Gaussian kernel, maps the data into an infinite-dimensional space using a Gaussian function. The sigmoid kernel applies the sigmoid function to transform the data.

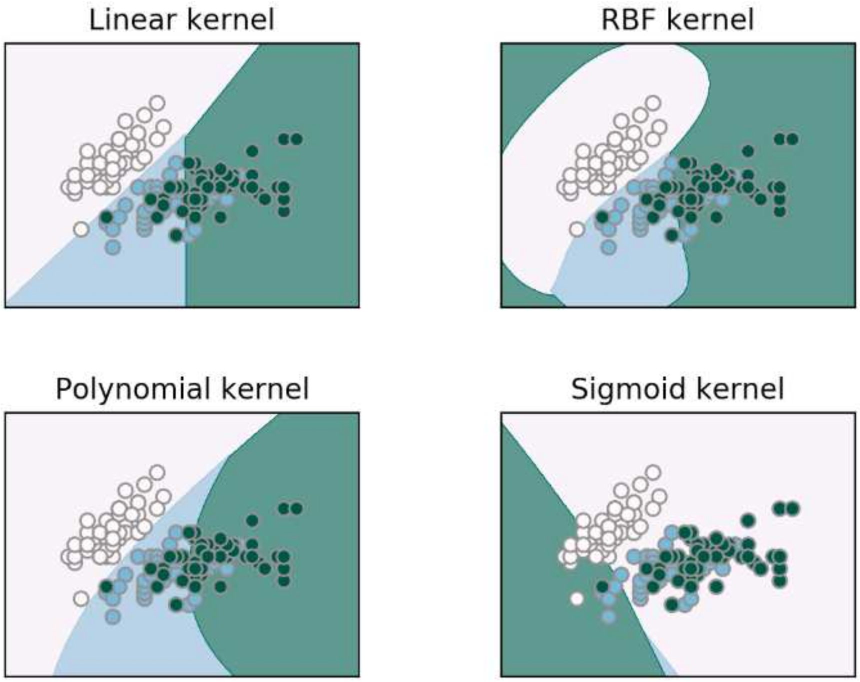


Figure 5.1: SVM Different Kernel Function Options

In this paper, we conducted a test on our data and explored the performance of the various kernel functions. Our objective was to determine the most effective kernel for accurately predicting categories within our labeled dataset. We identified the linear kernel as the optimal choice due to its superior performance in predicting the results.

5.3 Random Forest

The foundation of Random Forest can be attributed to the work on decision trees, which dates back to the 1960s and 1970s. The idea of using multiple decision trees for classification, known as "bagging" (Bootstrap Aggregating), was introduced by Breiman (1996). This work laid the groundwork for the ensemble approach that would later become the Random Forest. The term "Random Forest" and the specific algorithm that we know today were introduced by Breiman in a subsequent paper titled published in 2001.

Random Forest is a powerful and versatile ensemble machine learning algorithm that is widely used for both classification and regression tasks. It operates by constructing multiple decision trees during the training phase and combines their outputs to make predictions.

A decision tree is a graphical representation of a decision-making process or a classification system that resembles an upside-down tree. It consists of nodes and branches. It starts with a single node at the top, known as the root node, which represents the entire dataset. From this root node, branches extend to other nodes, which represent decisions or tests on specific features or attributes of the data. The final nodes, called leaf nodes, contain the outcome or the predicted class or value.

The process behind Random Forest is as follows: Randomly select N subsets of the training data (with replacement). Each subset is used to train a decision tree. For each subset, grow a decision tree by recursively splitting the data based on the most discriminative features. The Gini impurity or entropy may be used as splitting criteria. Repeat this process N times to build N decision trees. These trees form the Random Forest.

To make a prediction for a given text, pass it through each decision tree in the forest. Each decision tree independently assigns a class label (emotion) based on the text's features.

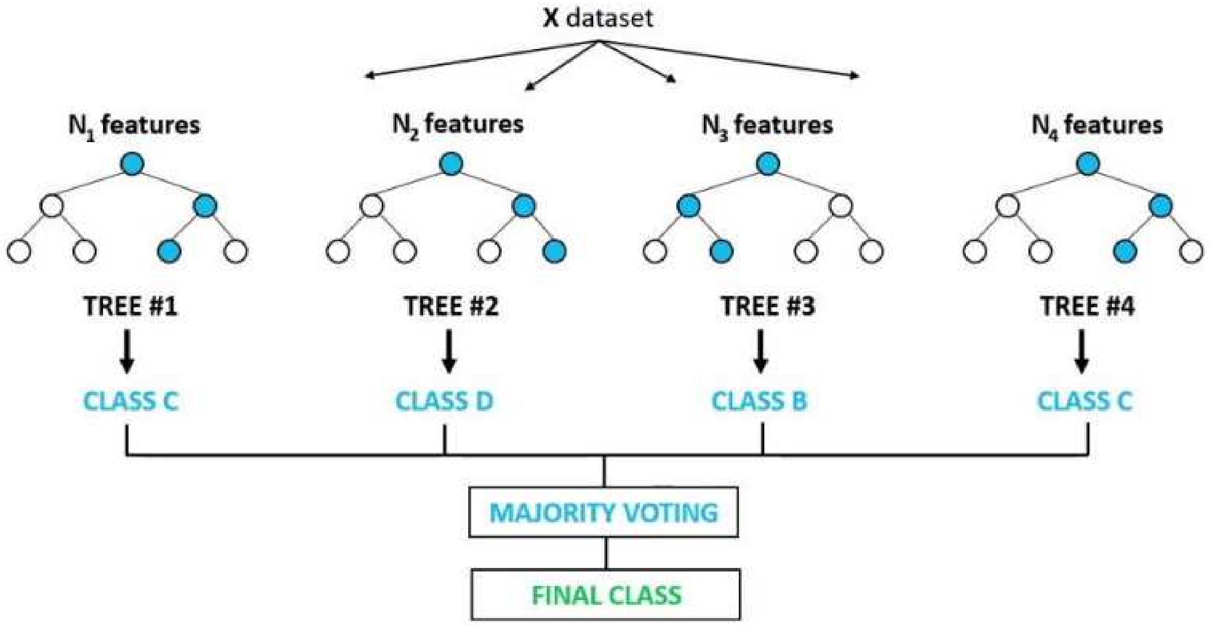


Figure 5.2: Random Forest Decision Process

These predictions can be aggregated by either:

Majority Voting: Each tree's vote contributes to the final prediction, and the class with the most votes is the predicted emotion.

Probability Averaging: Each tree provides a probability distribution over the classes, and the final prediction is based on the average probabilities.

It is worth nothing that for the purpose of this analysis, we conducted experiments with a variety of decision tree quantities within the Random Forest ensemble, specifically considering values such as 25, 50, 100, 150, 200, and 300. Subsequently, we assessed the outcomes associated with each of these hyperparameters and ultimately selected the one that demonstrated better performance at the prediction at hand. In our specific case, the optimal choice was a model featuring 100 decision trees.

5.4 Naive Bayes

The term "Bayesian" originates from Thomas Bayes, an 18th-century British statistician and theologian. His work laid the groundwork for Bayesian probability theory, which is at the core of the Naive Bayes classifier. Bayes' theorem, published posthumously in 1763, provided a way to update probability estimates based on new evidence, making it a fundamental concept in the field of probability and statistics.

However, the transition of Naive Bayes into a machine learning model for classification occurred in the mid-20th century, as computing power and data availability increased. One of the earliest papers that highlighted the use of Naive Bayes in classification was published by Duda et al. (1973). Their work demonstrated the effectiveness of Naive Bayes in distinguishing patterns and objects within a scene.

In the context of emotion recognition, Naive Bayes calculates the probability of an input text or data belonging to a particular emotional category. It does this by learning from labeled datasets where text or features are associated with specific emotions.

The formula for the probability that a given text belongs to a particular class is calculated using Bayes' theorem:

$$P(C|X) = \frac{P(X)P(C) \cdot P(X|C)}{P(X)}$$

Where $\mathcal{C}$ is the set of possible class labels. $\mathbf{X}$ is the feature vector representing the text. x_i represents the count of a specific word or token in the text. $P(C)$ is the prior probability of a class (the probability of a text belonging to a specific emotion without considering the text itself).

$P(X|C)$ is the conditional probability of observing the feature vector $\mathbf{X}$ given the class $\mathcal{C}$.

When using the Naive Bayes assumption, which assumes that the features are conditionally independent given the class, the formula simplifies to:

$$P(C|X) \propto P(C) \cdot \prod_{i=1}^{n} P(x_i|C)$$

In practice, you estimate $P(C)$ and $P(x_i|C)$ from the training data and classify the text into the class with the highest probability.

5.5 Results

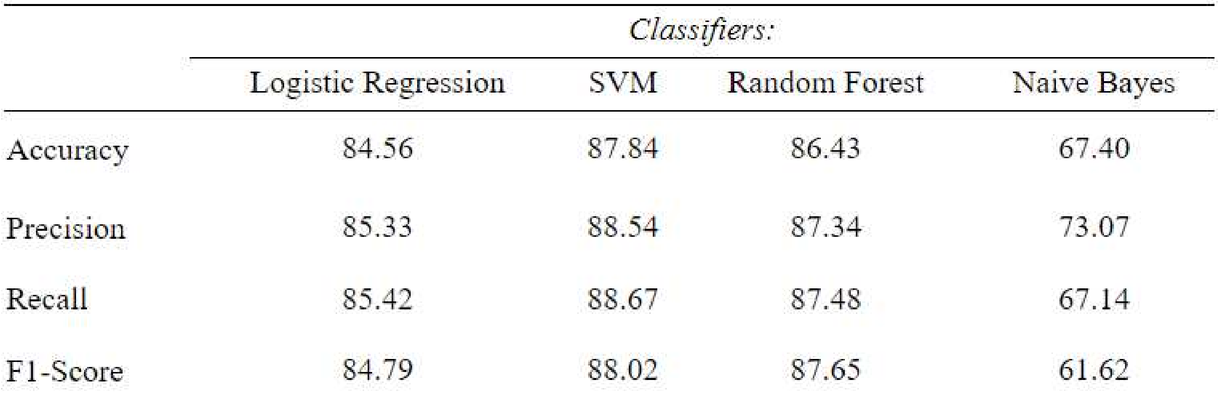
The performances of the classifiers were then compared in terms of accuracy, precision, recall and f1-measure. For all classifiers, a training set comprising 80% of the data was used for training, while the remaining 20% was reserved for testing and making predictions.

Accuracy measures the overall correctness of a classifier's predictions. It is the ratio of correctly classified instances to the total number of instances in the dataset. While accuracy provides a general sense of a classifier's performance, it can be misleading when dealing with imbalanced datasets where one class dominates.

Precision measures the accuracy of a classifier specifically for the positive class. It is the ratio of true positive predictions to the total number of positive predictions, and it quantifies the classifier's ability to avoid false positives. High precision indicates that the classifier makes fewer incorrect positive predictions.

Recall measures a classifier's ability to identify all positive instances. It is the ratio of true positive predictions to the total number of actual positive instances in the dataset. High recall suggests that the classifier captures most of the positive instances and avoids false negatives.

The F1-measure is the harmonic mean of precision and recall. It combines both precision and recall into a single score, which is particularly useful when you want a balanced measure of a classifier’s performance.

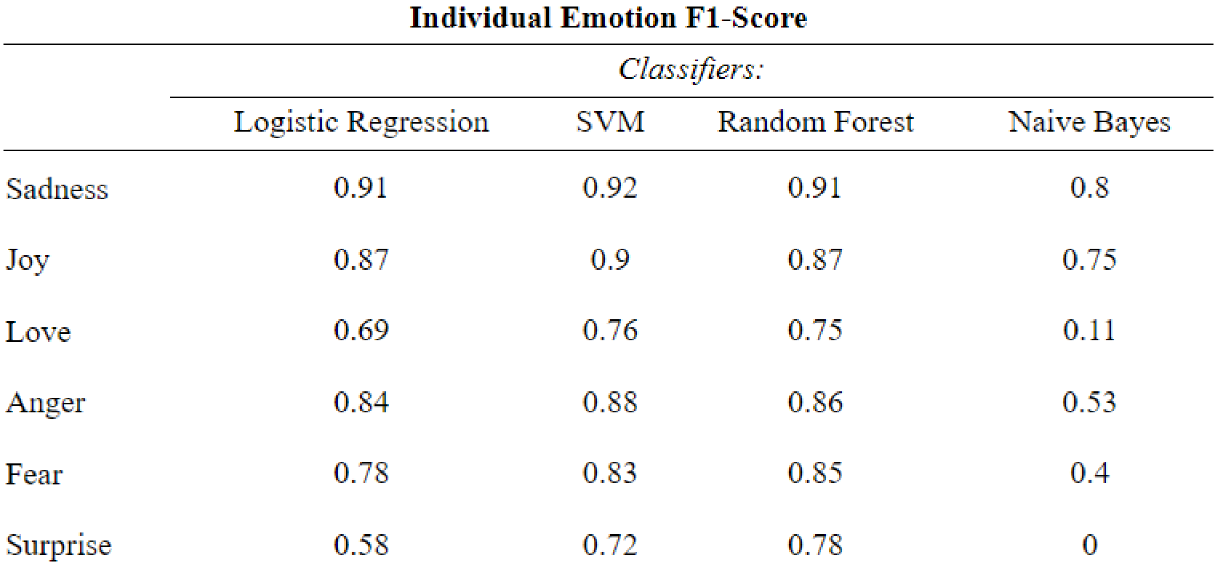


	Classifiers:			
	Logistic Regression	SVM	Random Forest	Naive Bayes
Accuracy	84.56	87.84	86.43	67.40
Precision	85.33	88.54	87.34	73.07
Recall	85.42	88.67	87.48	67.14
F1-Score	84.79	88.02	87.65	61.62

Figure 5.3: Results from the classifier (%)

In summary, each of the four classifiers was evaluated for its performance on the emotion classification task. SVM stands out as the top performer, offering high accuracy, precision, recall, and F1-score. Logistic Regression and Random Forest also provided reliable results, demonstrating balanced precision and recall. On the other hand, Naive Bayes, while offering good precision, exhibited lower recall and F1-score.

Next we will take a look at the individual F1-Score for each emotion in our group:



Individual Emotion F1-Score

	Classifiers:			
	Logistic Regression	SVM	Random Forest	Naive Bayes
Sadness	0.91	0.92	0.91	0.8
Joy	0.87	0.9	0.87	0.75
Love	0.69	0.76	0.75	0.11
Anger	0.84	0.88	0.86	0.53
Fear	0.78	0.83	0.85	0.4
Surprise	0.58	0.72	0.78	0

Figure 5.4: F1-Score for the emotions

Starting with the emotion "Sadness," all classifiers showcased strong performance. Logistic Regression, SVM, and Random Forest exhibited F1-Scores of 0.91, highlighting their efficacy in capturing the essence of sadness in the corpus. Meanwhile, Naive Bayes, with an F1-Score of 0.80 has a performance slightly below the others.

Looking at "Joy", the classifiers maintained their proficiency. Logistic Regression and Random Forest each achieved solid F1-Scores of 0.87. SVM excelled, with the highest F1-Score of 0.90, asserting itself as the top performer in this emotional category. Even Naive Bayes, with an F1-Score of 0.75, displayed a good performance.

However, when confronted with the emotion of "Love," classifiers presented varying results. SVM emerged as the leader with an F1-Score of 0.76, effectively identifying instances of love. In contrast, Logistic Regression and Random Forest, with F1-Scores of 0.69 and 0.75, respectively, maintained decent classification capabilities. Yet, Naive Bayes faced considerable challenges, manifesting in a notably lower F1-Score of 0.11, suggesting limitations in capturing the nuances of this particular emotion.

The emotion "Anger" brought about a solid performance from all classifiers, with SVM again attaining the highest F1-Score of 0.88, closely followed by Logistic Regression and Random Forest, which achieved F1-Scores of 0.84 and 0.86, respectively. Naive Bayes exhibited a slight lag behind the other classifiers with an F1-Score of 0.53.

In the context of "Fear," the classifiers delivered robust performance, with Random Forest emerging as the top performer, boasting an F1-Score of 0.85. Logistic Regression and SVM also secured solid F1-Scores of 0.78 and 0.83, respectively. Conversely, Naive Bayes faced difficulties in capturing expressions of fear, resulting in a lower F1-Score of 0.40.

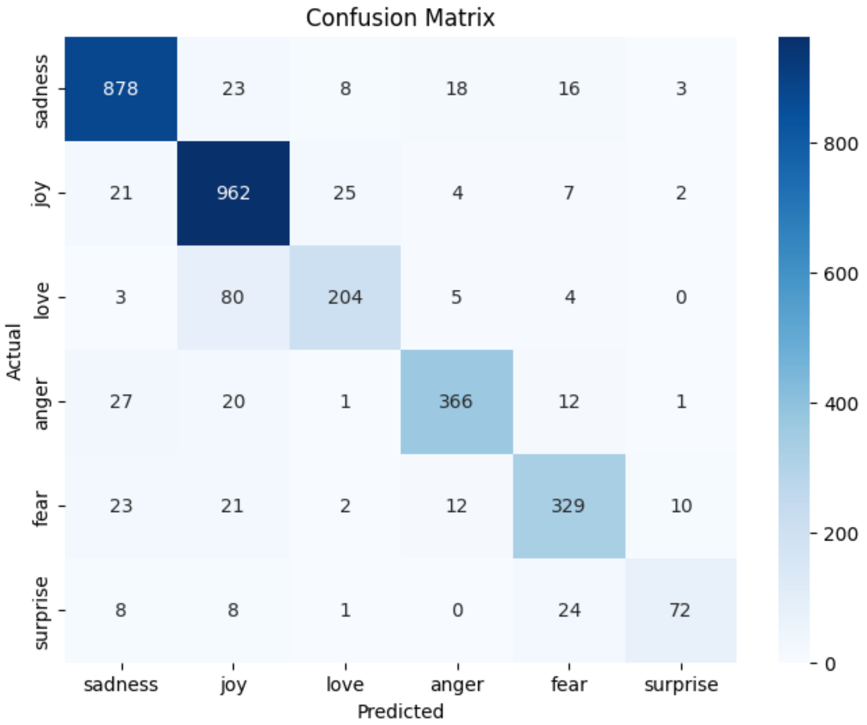
Lastly, for the emotion of "Surprise," the classifiers displayed varying performance. Random Forest excelled with the highest F1-Score of 0.78, indicating its effectiveness in recognizing instances of surprise. SVM also performed well with an F1-Score of 0.72. In contrast, Logistic

Regression and Naive Bayes faced challenges in capturing surprise-related content, with Logistic Regression yielding an F1-Score of 0.58 and Naive Bayes not being able to identify a single document for this emotion.

In summary, the F1-Scores for each emotion classification revealed the varying capabilities of the machine learning classifiers. SVM consistently demonstrated strong performance across multiple emotions, while Logistic Regression and Random Forest also delivered solid results in most cases. Naive Bayes, on the other hand, faced challenges in certain emotion classifications. These findings underscore the importance of selecting an appropriate classifier based on the specific emotional content analysis requirements and comprehending the strengths and limitations of each algorithm in different emotional contexts.

With all these results in mind, the SVM classifier was chosen for the subsequent application to unlabeled datasets encompassing Movies and Documentaries reviews. This decision will facilitate the prediction of underlying emotions associated with our reviews and serve as a pivotal component in modeling our outcomes.

We can take a further look at the data and examine the confusion matrix for the SVM classifier, which reveals the predicted outcomes in comparison to the actual labeled emotions (the true outcomes), it becomes apparent that the classifier exhibits higher confusion rates between emotions such as ’love’ and ’joy,’ as well as ’surprise’ and ’fear.’ This observation is further supported by the individual emotion F1-scores, where ’love’ (0.76) and ’surprise’ (0.72) received the lowest scores among all the emotions recognized.



	sadness	joy	love	anger	fear	surprise
sadness	878	23	8	18	16	3
joy	21	962	25	4	7	2
love	3	80	204	5	4	0
anger	27	20	1	366	12	1
fear	23	21	2	12	329	10
surprise	8	8	1	0	24	72

Figure 5.5: SVM Confusion Matrix

One plausible explanation for this outcome is the limited amount of labeled data available for both ’love’ and ’surprise.’ The classifier heavily relies on the training data it has access to, and when these emotions are underrepresented, it may struggle to discern them accurately. Consequently, this imbalance in the data distribution can lead to reduced performance in distinguishing ’love’ and ’surprise’ from other emotions in the classification process. However, the model still achieved a 0.76 and 0.72 F1-Score in ’love’ and ’surprise’ respectively. Which when dealing with categorical data is still considered a very good score.

You can find the confusion matrixes from the other classifiers used in the Annex I.

Chapter 6

Empirical Strategy

This paper aims to estimate the effect of the COVID-19 pandemic on the reviews ratings of scientific documentaries and the presence of an angry sentiment in these reviews. However, a naive comparison of the post-Pandemic ratings of scientific documentaries to those of other movies would not estimate a causal relationship, because of nherent differences in how the two groups of movies are rated. Similarly, analyzing a change in the ratings of scientific documentaries before and after the Pandemic, would also not yield a causal effect: because movie ratings might change with time, regardless of the genre.

We will use a Differences-in-Differences (DiD) framework, comparing the changes in ratings of scientific documentaries to that of other movies, before and after the pandemic. Scientific documentaries are designated as the treatment group and the top 500 movies dataset was selected as the control group.

This approach allows us to filter out time-invariant differences in the ratings of documentaries and other movies, as well as time-varying trends that affect all movies in the same manner. It thus establishes a causal effect of the Pandemic on the public's perception of scientific documentaries, provided that no time-varying differences exist in the ratings of scientific documentaries and other movies.

We estimate the following model:

$$y_{i,t} = \beta_0 + \beta_1 \text{Pandemic} + \beta_2 \text{Scientific} + \beta_3 (Pandemic * Scientific) + e_{i,t}$$

Here, the dependent variable $y_{i,t}$ corresponds to the quality of the reviews: either measured in ratings, or through a dummy for whether the review is flagged as containing angry emotions. In the equation, 'Pandemic' is a binary variable equal to 1 if the review was made during or after the pandemic, namely the years spanning from 2020 to 2022 and 0 otherwise. Therefore, β_1 represents the effect of a review being made during or after the Pandemic on $y_{i,t}$, compared to reviews made before this event. On the other hand, 'Scientific' is another binary variable, equal to 1 if the review pertains to a scientific documentary and 0 otherwise. β_2 thus measures the differences in $y_{i,t}$, between scientific documentaries and other movies.

The coefficient β_3 measures the effect of the interaction between the treatment status and post-treatment period. That is, how the gap in $y_{i,t}$ between scientific documentaries and other movies changed with the onset of the Pandemic.

The validity of the DiD approach hinges on the equal trends assumption, which posits that there are no time-varying differences between the treatment and control groups. To put it differently, without any intervention, both the treatment and control units would have maintained their parallel trends, and any unexpected events occurring in the post-treatment period would have affected them similarly.

While this assumption cannot be definitively proven, we can assess its validity through some key methods:

> Comparative Analysis of Pre-Intervention Changes: Analyze changes in the outcomes of the treatment and control groups in the periods leading up to the program's implementation. If the outcome trends exhibit parallel movement before the intervention's

commencement, it suggests that these trends would likely have continued similarly in the absence of the program.

Placebo Test Involving a Sham Treatment Group: Implement a placebo test by introducing a sham or fake treatment group that was not impacted by the Pandemic. A placebo test yielding a result of zero impact provides support for the equal-trend assumption.

Placebo Test Involving a Sham Outcome: Conduct a placebo test by using a simulated or fake outcome measure. A placebo test result of zero impact bolsters confidence in the equal-trend assumption.

Using an Artificial Treatment Date: Apply the difference-in-differences estimation method using a "fake" or artificial treatment date. This date is chosen in such a way that, in theory, it should have no real impact on the outcomes of interest.

These methods were employed to demonstrate the robustness of our model in the subsequent chapter, which presents the results.

6.1 Rating Model

Our primary outcome variable for this model is the review ‘Rating’ which represents the user rating of the documentaries and movies in our dataset. These ratings, as mentioned before, have values between 1 and 10. The coefficient β_3 will measure how the gap in ratings between scientific documentaries and other movies changed with the Pandemic, in absolute terms.

6.2 Emotion Model

For the second model, we will use 'Angry' as the outcome. We then apply the same DiD model as in the rating model, but because we have a binary dependent variable, all the coefficients now have to be interpreted as a change in the probability of having a review with angry sentiment.

The coefficient β_3 will measure how the gap in probability of having an angry review between scientific documentaries and other movies changed with the Pandemic. Thus allowing us to examine whether the pandemic has heightened individuals' expressions of anger towards scientific content and whether this phenomenon is reflected in the reviews.

Chapter 7

Results

7.1 Testing the Parallel Trend Assumption

The DiD method hinges on the assumption that, in the absence of treatment, the treated and control groups would follow parallel trends over time. This assumption is fundamental for making causal claims about the treatment effect.

One way to assess the parallel trend assumption is through visual inspection of the pre-treatment period data. Plotting the outcome variable for both groups over time can reveal any noticeable deviations in trends. Any abrupt changes or divergences in the trajectories may suggest issues with the assumption.

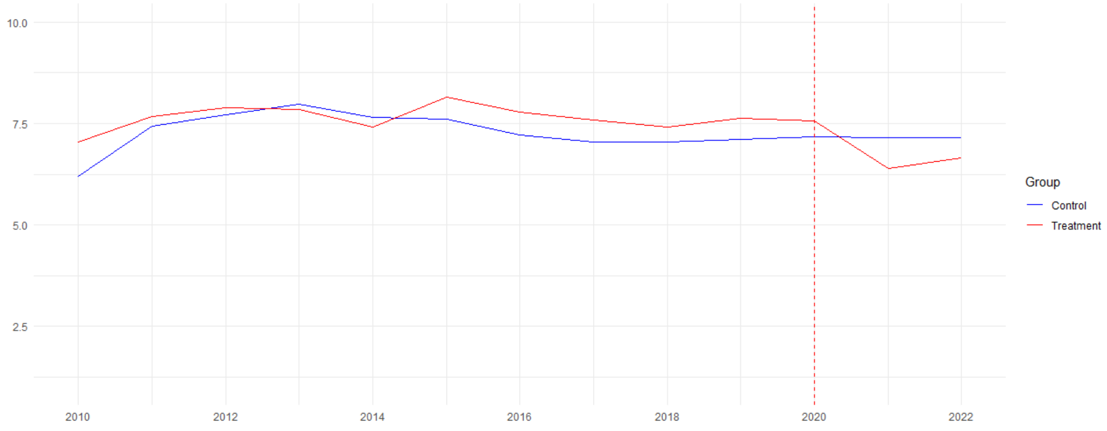


Figure 7.1: Parallel Trends - 12 years on Rating

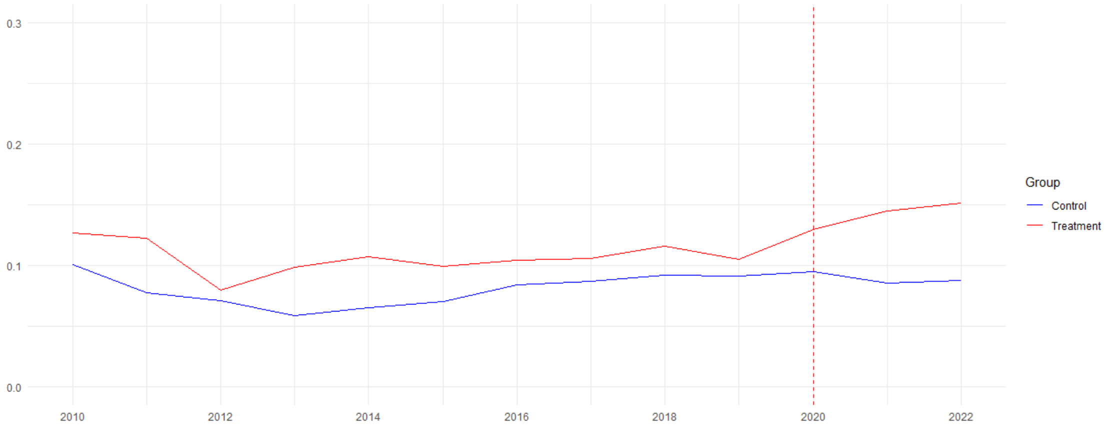


Figure 7.2: Parallel Trends - 12 years on Emotion (Angry)

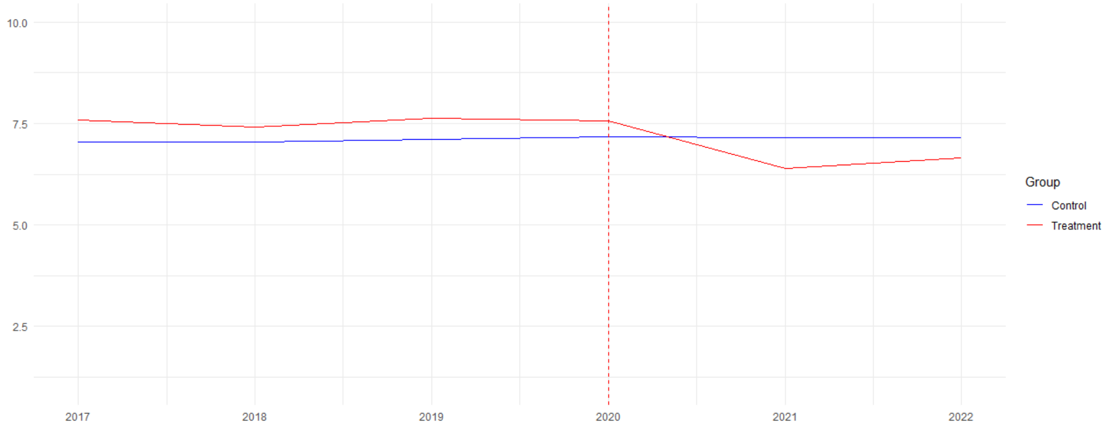


Figure 7.3: Parallel Trends - 6 years on Rating

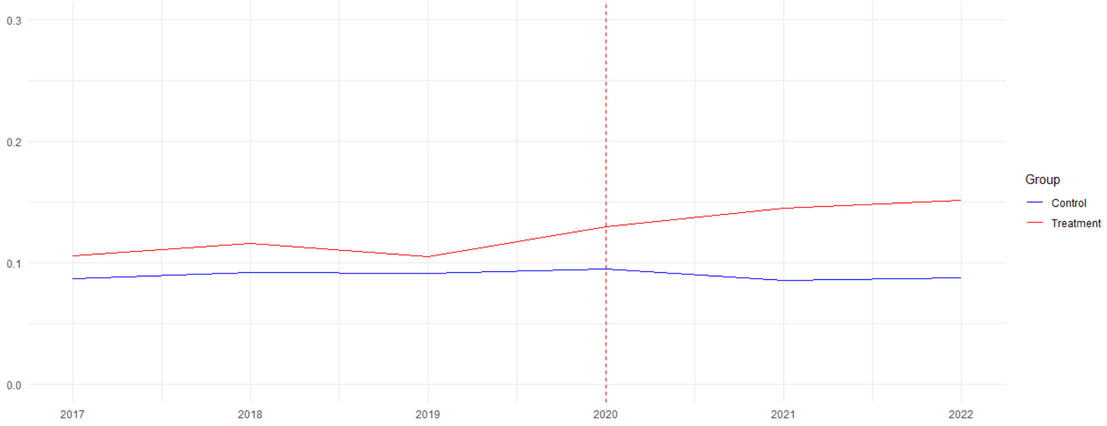


Figure 7.4: Parallel Trends - 6 years on Emotion (Angry)

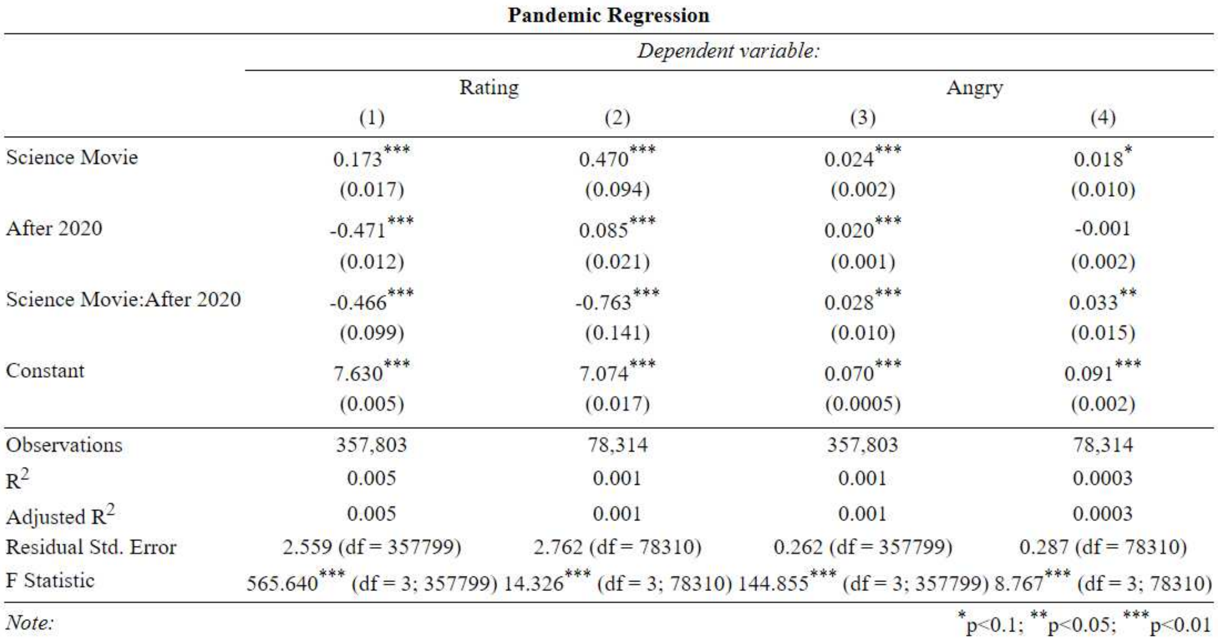
The figures presented above offer valuable insights into the temporal trends of both movie and documentary ratings. Prior to the year 2020, the ratings for both movies and documentaries exhibited similar trajectories, suggestive of parallel trends. However, post-2020, a notable deviation becomes evident. In particular, the ratings for scientific documentaries show a pronounced decline of more than one point on the rating scale.

Turning our attention to the graphs depicting the expression of anger, we again observe parallel trends leading up to 2020. However, a sudden surge in the frequency of angry emotional expressions becomes apparent within the treatment group after the year 2020. This sharp increase may signify a change in the emotional responses of individuals to certain stimuli or circumstances.

7.2 Differences-in-Differences Results

The regression results represent a comprehensive analysis of the impact of the pandemic on scientific documentaries reception and the resulting audience response, conducted through different models:

Models 1 and 2 focus on evaluating the effects on movie ratings, whereas models 3 and 4 examine the occurrence of angry sentiment within the reviews. In terms of time periods, it's important to note that 2020 serves as the initial 'post-treatment' year for all models, marking the onset of the pandemic's impact. Models 2 and 4 encompass a specific time frame, spanning from 2017 to 2022, which encapsulates three years prior to the pandemic and three years afterward. On the other hand, Models 1 and 3 cover the entire dataset period, extending from 2010 to 2022, providing a comprehensive perspective on the long-term impact assessment.



Pandemic Regression

	Dependent variable:			
	Rating		Angry	
	(1)	(2)	(3)	(4)
Science Movie	0.173^{***}	0.470^{***}	0.024^{***}	0.018^{*}
	(0.017)	(0.094)	(0.002)	(0.010)
After 2020	-0.471^{***}	0.085^{***}	0.020^{***}	-0.001
	(0.012)	(0.021)	(0.001)	(0.002)
Science Movie:After 2020	-0.466^{***}	-0.763^{***}	0.028^{***}	0.033^{**}
	(0.099)	(0.141)	(0.010)	(0.015)
Constant	7.630^{***}	7.074^{***}	0.070^{***}	0.091^{***}
	(0.005)	(0.017)	(0.0005)	(0.002)
Observations	357,803	78,314	357,803	78,314
R^2	0.005	0.001	0.001	0.0003
Adjusted R^2	0.005	0.001	0.001	0.0003
Residual Std. Error	2.559 (df = 357799)	2.762 (df = 78310)	0.262 (df = 357799)	0.287 (df = 78310)
F Statistic	565.640^{***} (df = 3; 357799)	14.326^{***} (df = 3; 78310)	144.855^{***} (df = 3; 357799)	8.767^{***} (df = 3; 78310)
Note:				*p<0.1; **p<0.05; ***p<0.01

Figure 7.5: DiD Regression for the 4 models

Several noteworthy observations can be made from the regression coefficients:

Science Movie Effect: The coefficient for "Science Movie" in both Model 1 and Model 2 shows a positive impact on ratings, signifying that viewers tend to rate scientific documentaries higher than other movies. However, in Model 3 and Model 4, this variable indicates a positive effect on the probability of evoking anger in reviewers.

After 2020 Effect: The "After 2020" variable has a negative coefficient in model 1 what would indicate an adverse impact of the pandemic on movie ratings, however in model 2, when looking at a shorter time period we see that there is actually an increase after the pandemic, although by a small figure. In model 3 the Pandemic seems to have generated an increased likelihood of reviews containing expressions of anger, and in model 4 the coefficient has no significance.

Constant: The constant term in all models represents the baseline rating or likelihood of anger in reviews in the pre-pandemic period. It's significantly different for each model, showcasing the nuances in baseline audience sentiment or ratings for different time periods and outcomes.

Interaction Effect: The "Science Movie:After 2020" interaction term, which is the term of interest for this paper, in both Model 1 and Model 2 displays a negative coefficient, suggesting that the negative effect of the pandemic on ratings is stronger for scientific documentaries. The coefficients in both models are significant at the 1% level.

The same interaction term in Model 3 and Model 4 indicates a positive effect, implying that the increase in the likelihood of a review being angry during or after the pandemic was stronger for scientific documentaries. Again, the coefficients in both models are significant at the 1% and 5% level, respectively.

In summary, these regression results provide valuable insights into the changing landscape of audience responses to scientific documentaries during and after the pandemic. They demonstrate not only the pandemic's negative impact on scientific documentaries ratings, but also its potential to elicit anger in audience reviews.

7.3 Robustness

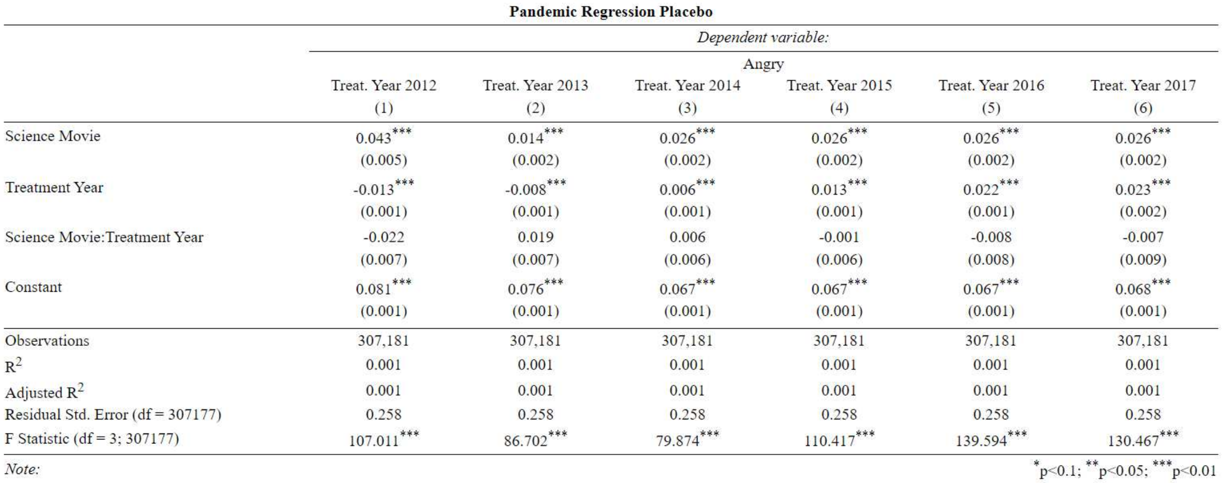
7.3.1 Placebo Tests

Using a placebo test in a DiD model is a statistical technique employed to assess the robustness of the main findings and to help ensure that the estimated treatment effect is not due to random noise or other confounding factors.

We introduced three placebos tests. In the first placebo test, we deliberately selected different artificial periods as potential "Treatment" dates, with the expectation that the coefficient of interest should not exhibit statistical significance under these conditions.

To select these artificial treatment periods, we followed a specific criterion. We only considered years that had at least two years of data both before and after the treatment date, mirroring the conditions of the original model. Accordingly, we identified the years 2012, 2013, 2014, 2015, 2016, and 2017 as suitable candidates for the placebo tests. Each of these years is individually represented in models 1 through 6, respectively.

We then use the exact same Angry model used before to estimate the coefficients of these different artificial treatment years:



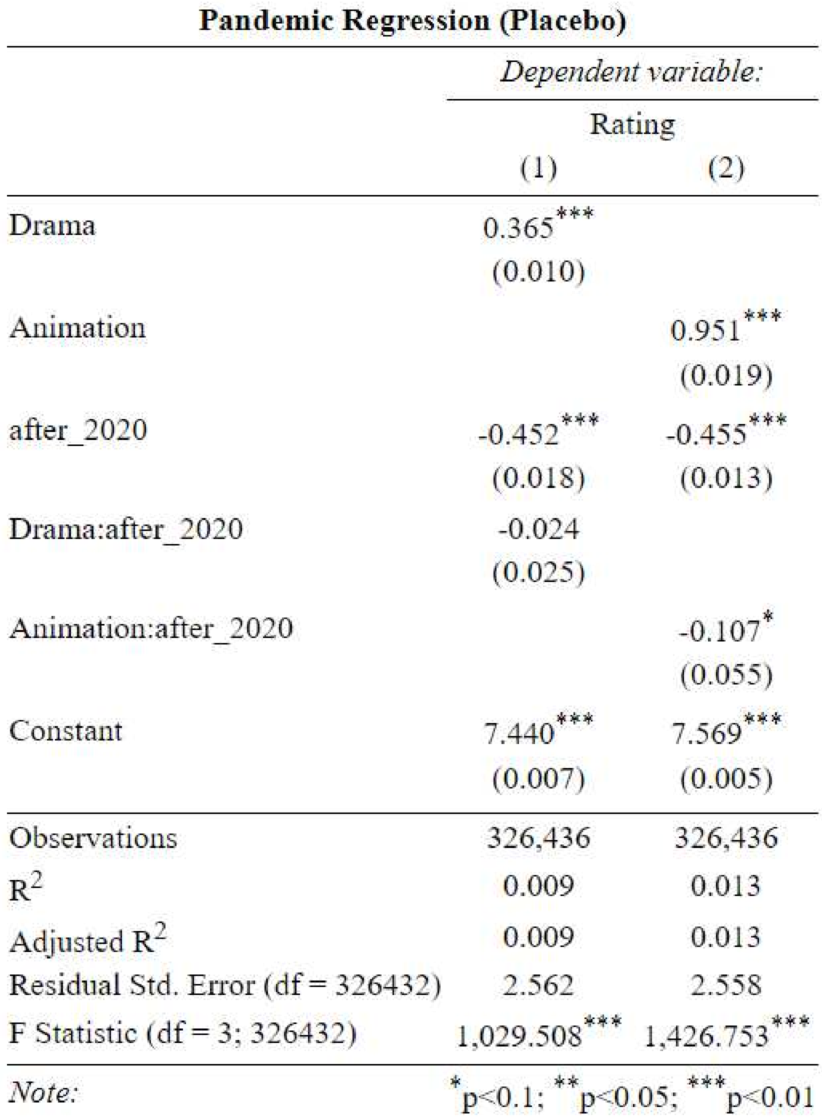
Pandemic Regression Placebo

	Dependent variable:					
	Angry					
	Treat. Year 2012	Treat. Year 2013	Treat. Year 2014	Treat. Year 2015	Treat. Year 2016	Treat. Year 2017
	(1)	(2)	(3)	(4)	(5)	(6)
Science Movie	0.043***	0.014***	0.026***	0.026***	0.026***	0.026***
	(0.005)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Treatment Year	-0.013***	-0.008***	0.006***	0.013***	0.022***	0.023***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)
Science Movie:Treatment Year	-0.022	0.019	0.006	-0.001	-0.008	-0.007
	(0.007)	(0.007)	(0.006)	(0.006)	(0.008)	(0.009)
Constant	0.081***	0.076***	0.067***	0.067***	0.067***	0.068***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Observations	307,181	307,181	307,181	307,181	307,181	307,181
R^2	0.001	0.001	0.001	0.001	0.001	0.001
Adjusted R^2	0.001	0.001	0.001	0.001	0.001	0.001
Residual Std. Error (df = 307177)	0.258	0.258	0.258	0.258	0.258	0.258
F Statistic (df = 3; 307177)	107.011***	86.702***	79.874***	110.417***	139.594***	130.467***
Note:						*p<0.1; **p<0.05; ***p<0.01

Figure 7.6: DiD Regression - Artificial Treatment Dates

The statistical analysis indicates that there is no significant effect resulting from the interaction between the treatment status and the post-treatment period. All the graphs of the parallel trends of this placebo test can are reported in Figures 9.4 to 9.9 in Annex 1.

Additionally, for the second test, we used other movie genres instead of science documentaries for our treatment group. We have chosen 'Drama' and 'Animation' genres as substitutes.



Pandemic Regression (Placebo)

	Dependent variable:	
	Rating	
	(1)	(2)
Drama	0.365***	
	(0.010)	
Animation		0.951***
		(0.019)
after_2020	-0.452***	-0.455***
	(0.018)	(0.013)
Drama:after_2020	-0.024	
	(0.025)	
Animation:after_2020		-0.107*
		(0.055)
Constant	7.440***	7.569***
	(0.007)	(0.005)
Observations	326,436	326,436
R^2	0.009	0.013
Adjusted R^2	0.009	0.013
Residual Std. Error (df = 326432)	2.562	2.558
F Statistic (df = 3; 326432)	1,029.508***	1,426.753***
Note:	*p<0.1; **p<0.05; ***p<0.01	

Figure 7.7: DiD Regression - Family and Animation

We see that there is no significant effect resulting from the interaction between the treatment status and the post-treatment period for the 'Drama' genre. In 'Animation' we actually find a significance level of $p < 0.1$. This indicates a degree of significance, albeit at a relatively modest level.

We then proceed to examine the parallel trends assumption for each one:

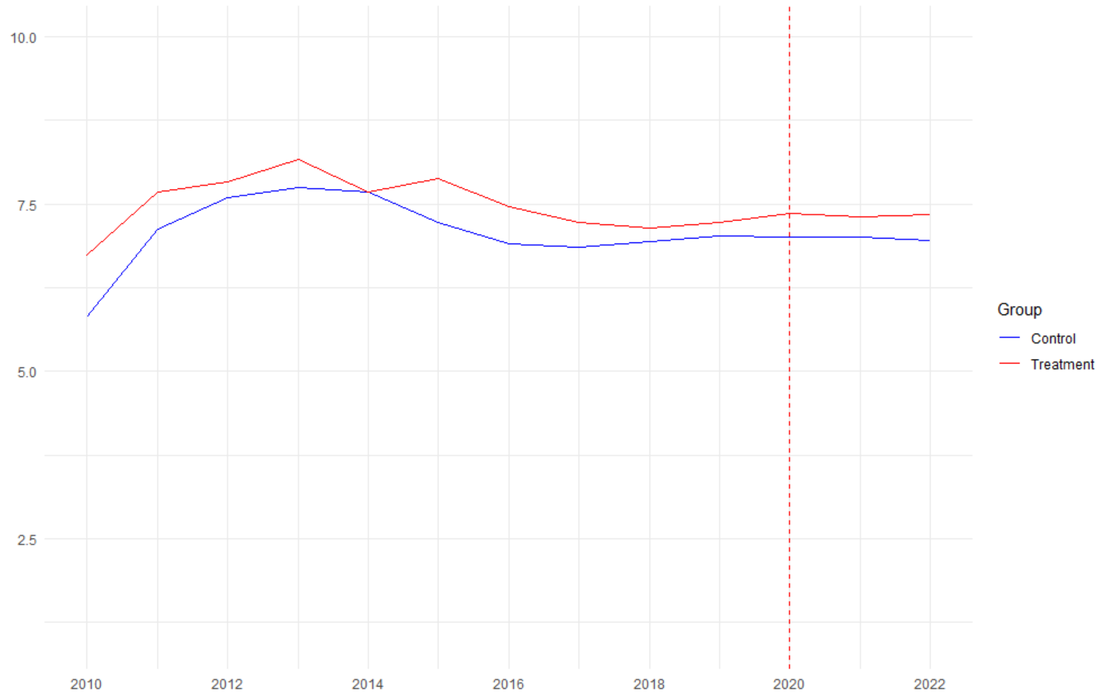


Figure 7.8: Parallel Trends - 12 years on Rating for Drama movies

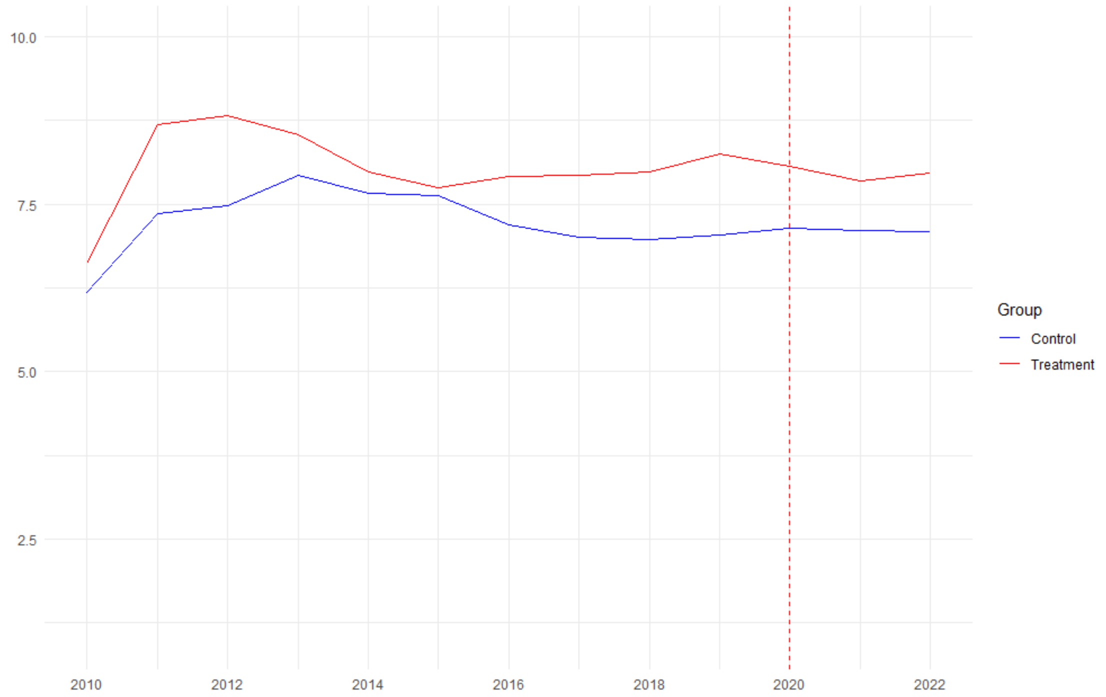
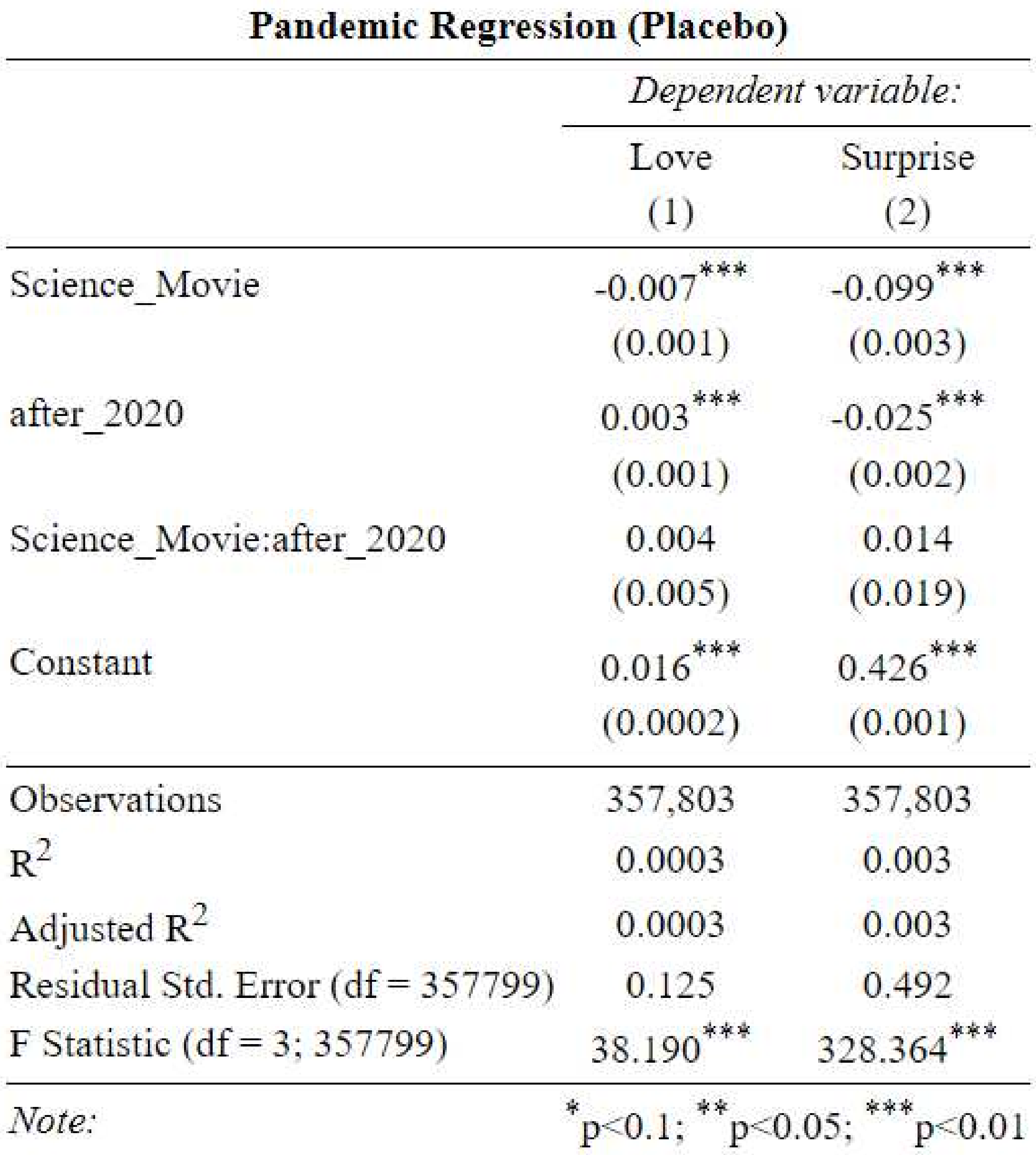


Figure 7.9: Parallel Trends - 12 years on Rating for Animation movies

When evaluating the 'Drama' and 'Animation' genres as the treatment groups, we can also observe that the parallel trends assumption holds for both genres.

In the third test, we selected two different sentiment outcomes that, in theory, should not be affected by the treatment we are studying. If our main DiD result is indeed capturing a causal effect, the placebo test should show no significant impact.In our placebo analysis, we opted to examine the impact of two distinct emotional states, namely, love and surprise, as our outcome variables.



Pandemic Regression (Placebo)

	Dependent variable:	
	Love	Surprise
	(1)	(2)
Science_Movie	-0.007^{***}	-0.099^{***}
	(0.001)	(0.003)
after_2020	0.003^{***}	-0.025^{***}
	(0.001)	(0.002)
Science_Movie:after_2020	0.004	0.014
	(0.005)	(0.019)
Constant	0.016^{***}	0.426^{***}
	(0.0002)	(0.001)
Observations	357,803	357,803
R^2	0.0003	0.003
Adjusted R^2	0.0003	0.003
Residual Std. Error (df = 357799)	0.125	0.492
F Statistic (df = 3; 357799)	38.190^{***}	328.364^{***}
Note:	*p<0.1; **p<0.05; ***p<0.01	

Figure 7.10: DiD Regression on Love and Surprise Outcomes

The statistical analysis indicates that there is no significant effect resulting from the interaction between the treatment status and the post-treatment period.

By conducting these placebos tests, we have addressed several concerns:

Confounding Factors: The placebo test helps ensure that our treatment effect isn't driven by unobserved confounding factors that may have affected both the treatment and control groups.

Random Fluctuations: It helps to rule out the possibility that the treatment effect is merely due to random fluctuations or noise in the data.

Causal Inference: It strengthens your case for causal inference by providing evidence that the treatment caused the observed effect.

7.3.2 Augmented Dickey-Fuller (ADF)

The Augmented Dickey-Fuller (ADF) test is used to determine whether a time series is stationary or contains a unit root (non-stationary). It is one of the most commonly used statistical test when it comes to analyzing the stationary of a series.

First, we should understand the Dickey-Fuller Test. It uses an autoregressive model and optimizes an information criterion across multiple different lag values.

It uses the Null Hypothesis (H_0): $\alpha = 1$

The DF equation follows as:

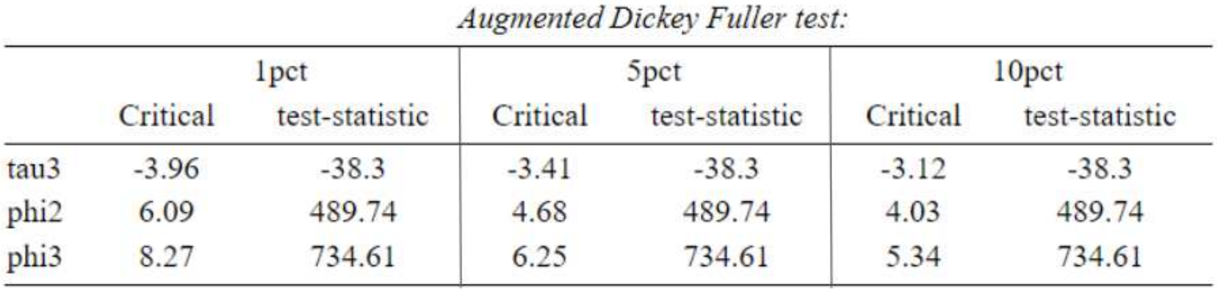
$$y_t = c + \beta t + \alpha y_{t-1} + \epsilon_t$$

The ‘augmented’ version of the Dickey-Fuller test expands the Dickey-Fuller test equation to include a high-order regressive process in the model.

$$y_t = c + \beta t + \alpha y_{t-1} + \Phi_1 \Delta Y_{t-1} + \Phi_2 \Delta Y_{t-2} + \ldots + \Phi_p \Delta Y_{t-p} + \epsilon_t$$

In the new equation we have simply added more differencing terms, the rest of the equation remains the exact same as in the DF test.

Using the ADF test in our data, we get the following results:



Augmented Dickey Fuller test:

	1pct		5pct		10pct	
	Critical	test-statistic	Critical	test-statistic	Critical	test-statistic
tau3	-3.96	-38.3	-3.41	-38.3	-3.12	-38.3
phi2	6.09	489.74	4.68	489.74	4.03	489.74
phi3	8.27	734.61	6.25	734.61	5.34	734.61

Figure 7.11: ADF Results

Value of test-statistic: The ADF test statistic are -38.33, 489.74 and 734.61, and it has corresponding critical values provided below.

Critical values for test statistics: These are the critical values that the ADF test statistic is compared against to make a decision regarding stationarity. Based on the results, the test statistic is much smaller than the critical values for all the given confidence levels (1%, 5%, and 10%), which suggests that the data is stationary. The ADF test rejects the null hypothesis of non-stationarity.

In summary, the ADF test suggests that the data is stationary, as the test statistic is significantly smaller than the critical values, and the model estimated is statistically significant with highly significant coefficients. The complete ADF test results including coefficients results can be found in Annex I.

Chapter 8

Conclusion

The COVID-19 pandemic has had a significant impact on multiple aspects of life. This study aims to examine how this unparalleled global event has affected public opinions on science. In order to do this, we analyze the pandemic's influence on scientific documentary ratings and emotional responses.

To estimate this effect, a Differences-in-Differences methodology was employed. The research design consisted of two primary models: one evaluating changes in ratings and the other one analyzing shifts in emotional responses, specifically expressions of anger in reviews.

The findings demonstrate several notable patterns. First, scientific documentaries generally received higher ratings than other film genres, suggesting audience appreciation for science-related content. Second, the pandemic years are associated with a decrease in ratings in movies of all genres.

Third, relative to other types of movies, the negative effect of the Pandemic on ratings of scientific documentaries was stronger. This finding indicates that while audiences may have had an initially favorable view of scientific documentaries, the pandemic has somehow increased their likelihood to express dissatisfaction or frustration with this genre.

It is estimated that the ratings of scientific documentaries decreased by an additional 0.47 compared to other movies after the pandemic. This is a significant effect, given that ratings range from 1 to 10. When looking at a narrower timeframe of 2017 to 2022, the decrease is even more significant, at approximately 0.76.

When looking at sentiments instead of ratings, the results are similar. The Pandemic had a more intense effect on the likelihood of angry reviews for scientific documentaries, which was increased by an additional 0.028 - 0.033 in comparison to other genres.

Our findings are supported by the placebo tests conducted in this research, increasing their veracity. By testing fabricated treatment dates, various film genres that are not likely to be affected by the pandemic and other sentiment outcomes that not 'angry', we demonstrate that the difference-in-differences (DiD) results notably detect a causal impact unique to scientific documentaries amid the pandemic season. This reinforces the idea that alterations in ratings and emotional responses caused by the pandemic are not random, but instead closely tied to the subject matter of scientific documentaries.

Bibliography

[1] Christos-Nikolaos Anagnostopoulos, Theodoros Iliou, and Ioannis Giannoukos. Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011. *Artificial Intelligence Review*, 43:155–177, 2015.

[2] Alexandra Balahur, Jesus M Hermida, and Andres Montoyo. Building and exploiting emotinet, a knowledge base for emotion detection based on the appraisal theory model. *IEEE transactions on affective computing*, 3(1):88–101, 2011.

[3] Leo Breiman. Bagging predictors. *Machine learning*, 24:123–140, 1996.

[4] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20:273–297, 1995.

[5] Richard O Duda, Peter E Hart, et al. *Pattern classification and scene analysis*, volume 3. Wiley New York, 1973.

[6] Paul Ekman. Facial expression and emotion. *American psychologist*, 48(4):384, 1993.

[7] Despina Gavresi and Anastasia Litina. Past exposure to macroeconomic shocks and populist attitudes in europe. *Journal of Comparative Economics*, 2023.

[8] Paul Gill, Kate Stewart, Elizabeth Treasure, and Barbara Chadwick. Methods of data collection in qualitative research: interviews and focus groups. *British dental journal*, 204(6):291–295, 2008.

[9] Maryam Hasan, Elke Rundensteiner, and Emmanuel Agu. Emotex: Detecting emotions in twitter messages. 2014.

[10] Dou Hu, Lingwei Wei, and Xiaoyong Huai. Dialoguecrn: Contextual reasoning networks for emotion recognition in conversations. *arXiv preprint arXiv:2106.01978*, 2021.

[11] Theodoros Iliou and Christos-Nikolaos Anagnostopoulos. Comparison of different classifiers for emotion recognition. In *2009 13th Panhellenic Conference on Informatics*, pages 102–106. IEEE, 2009.

[12] Filip Kostelka. Does democratic consolidation lead to a decline in voter turnout? global evidence since 1939. *American Political Science Review*, 111(4):653–667, 2017.

[13] Filip Kostelka. The state of political participation in post-communist democracies: Low but surprisingly little biased citizen engagement. In *The State of Democracy in Central and Eastern Europe*, pages 105–128. Routledge, 2017.

[14] Sarah E Kreps and Douglas L Kriner. Model uncertainty, political contestation, and public trust in science: Evidence from the covid-19 pandemic. *Science advances*, 6(43):eabd4563, 2020.

[15] Chi-Chun Lee, Emily Mower, Carlos Busso, Sungbok Lee, and Shrikanth Narayanan. Emotion recognition using a hierarchical binary decision tree approach. *Speech Communication*, 53(9-10):1162–1171, 2011.

[16] Yi-Lin Lin and Gang Wei. Speech emotion recognition based on hmm and svm. In *2005 international conference on machine learning and cybernetics*, volume 8, pages 4898–4901. IEEE, 2005.

[17] Niels G Mede and Mike S Schäfer. Science-related populism: Conceptualizing populist demands toward science. *Public Understanding of science*, 29(5):473–491, 2020.

[18] Kamoltep Moolthaisong and Wararat Songpan. Emotion analysis and classification of movie reviews using data mining. In *2020 international conference on data science, artificial intelligence, and business analytics (DATABIA)*, pages 89–92. IEEE, 2020.

[19] Myriam Munezero, Calkin Suero Montero, Tuomo Kakkonen, Erkki Sutinen, Maxim Mozgovoy, and Vitaly Klyuev. Automatic detection of antisocial behaviour in texts. *Informatica*, 38(1), 2014.

[20] Myriam Douce Munezero. *Leveraging emotion and word-based features for antisocial behavior detection in user-generated content*. PhD thesis, Itä-Suomen yliopisto, 2017.

[21] Robert Plutchik. The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American scientist*, 89(4):344–350, 2001.

[22] Ashwin Rao, Fred Morstatter, Minda Hu, Emily Chen, Keith Burghardt, Emilio Ferrara, and Kristina Lerman. Political partisanship and antiscience attitudes in online discussions about covid-19: Twitter content analysis. *Journal of medical Internet research*, 23(6):e26692, 2021.

[23] Jonathan Roth, Pedro HC Sant'Anna, Alyssa Bilinski, and John Poe. What's trending in difference-in-differences? a synthesis of the recent econometrics literature. *Journal of Econometrics*, 2023.

[24] Viktor Rozgić, Sankaranarayanan Ananthakrishnan, Shirin Saleem, Rohit Kumar, and Rohit Prasad. Ensemble of svm trees for multimodal emotion recognition. In *Proceedings of the 2012 Asia Pacific signal and information processing association annual summit and conference*, pages 1–4. IEEE, 2012.

[25] Elvis Saravia, Hsien-Chi Toby Liu, Yen-Hao Huang, Junlin Wu, and Yi-Shin Chen. Carer: Contextualized affect representations for emotion recognition. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 3687–3697, 2018.

[26] Carlo Strapparava and Rada Mihalcea. Learning to identify emotions in text. In *Proceedings of the 2008 ACM symposium on Applied computing*, pages 1556–1560, 2008.

[27] Usman Tariq, Kai-Hsiang Lin, Zhen Li, Xi Zhou, Zhaowen Wang, Vuong Le, Thomas S Huang, Xutao Lv, and Tony X Han. Emotion recognition from an ensemble of features. In *2011 IEEE International Conference on Automatic Face & Gesture Recognition (FG)*, pages 872–877. IEEE, 2011.

Chapter 9

Annex I

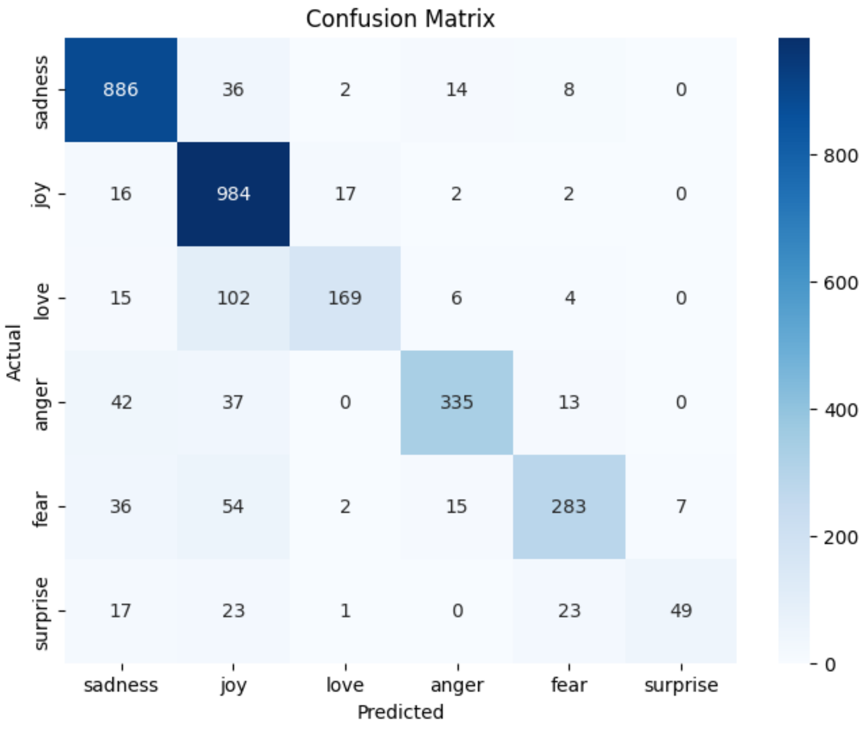


Figure 9.1: Logistic Regression Confusion Matrix

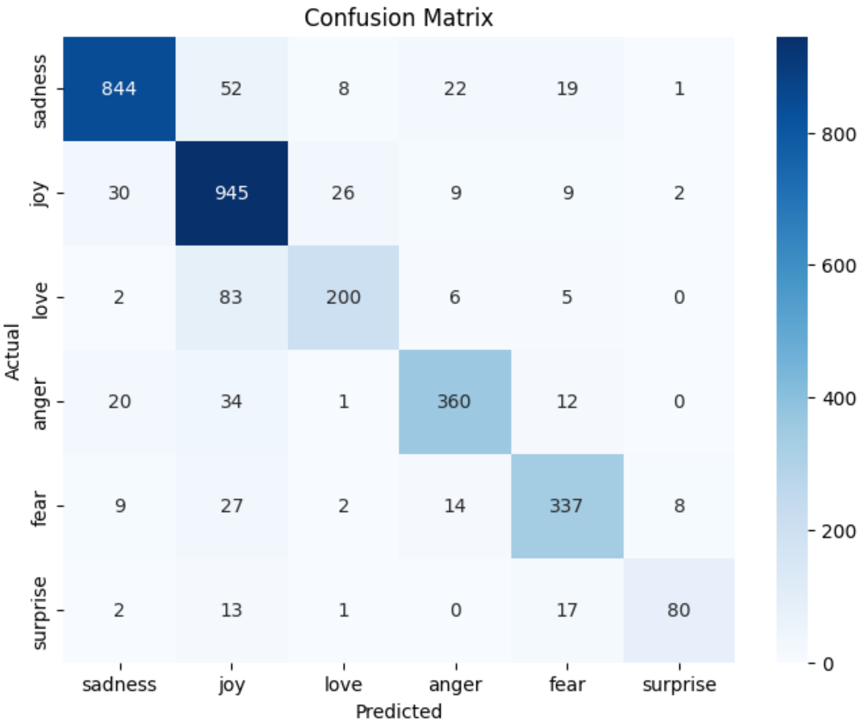


Figure 9.2: Random Forest Confusion Matrix

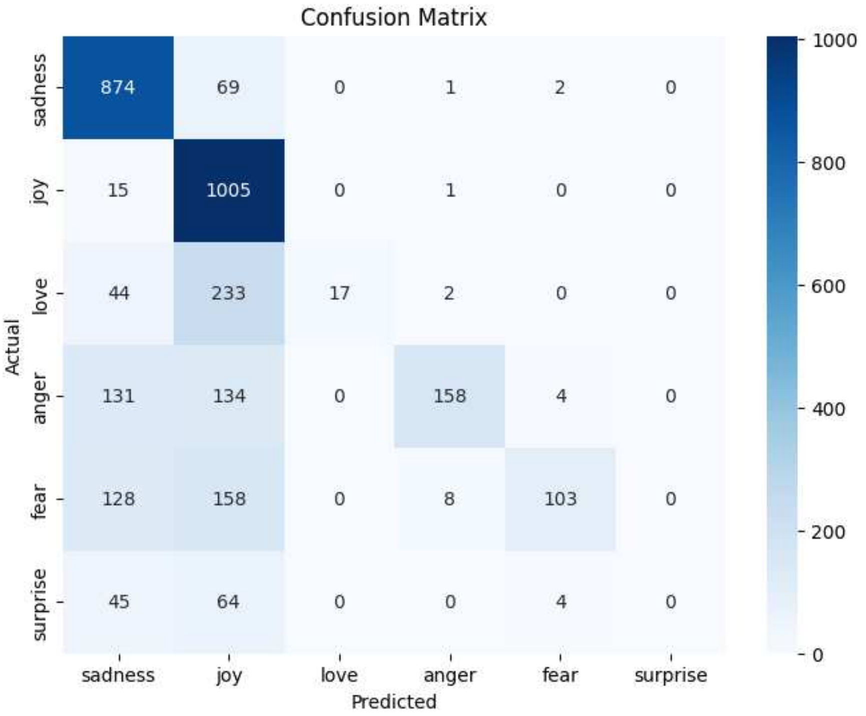


Figure 9.3: Naive Bayes Confusion Matrix

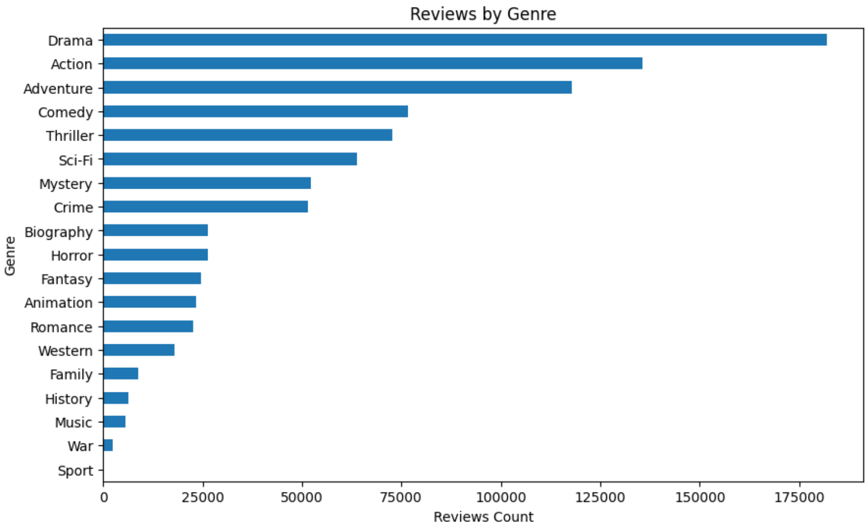


Figure 9.4: Top 500 Dataset - Reviews by Genre

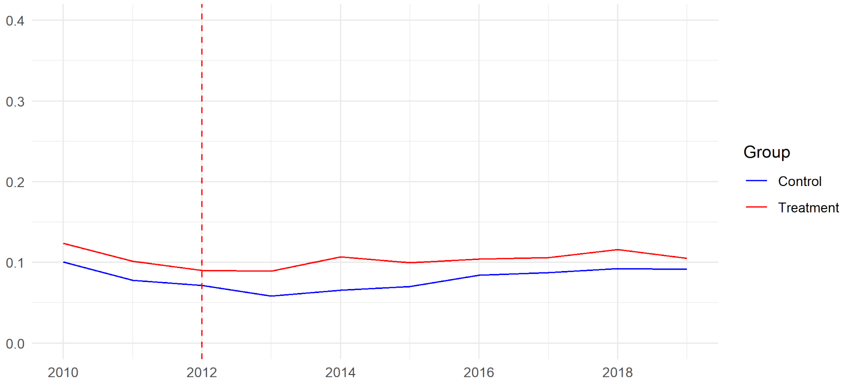


Figure 9.5: Parallel Trends - 12 years on Emotion (Angry) for 2012 as treatment

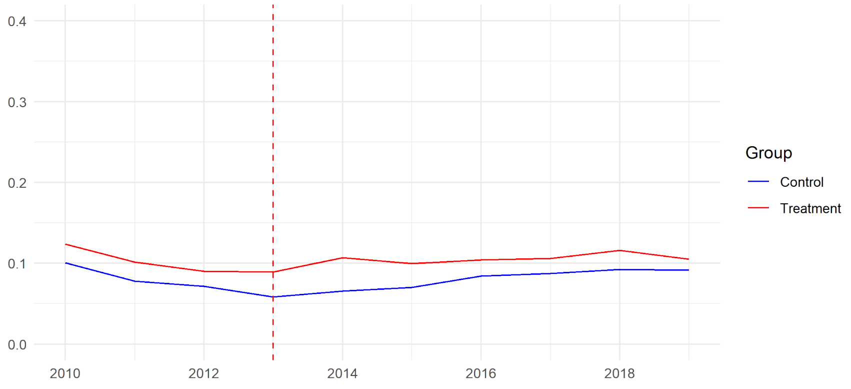


Figure 9.6: Parallel Trends - 12 years on Emotion (Angry) for 2013 as treatment

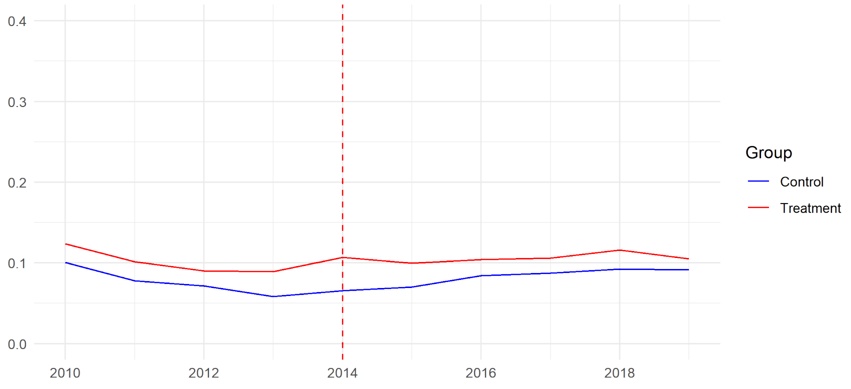


Figure 9.7: Parallel Trends - 12 years on Emotion (Angry) for 2014 as treatment

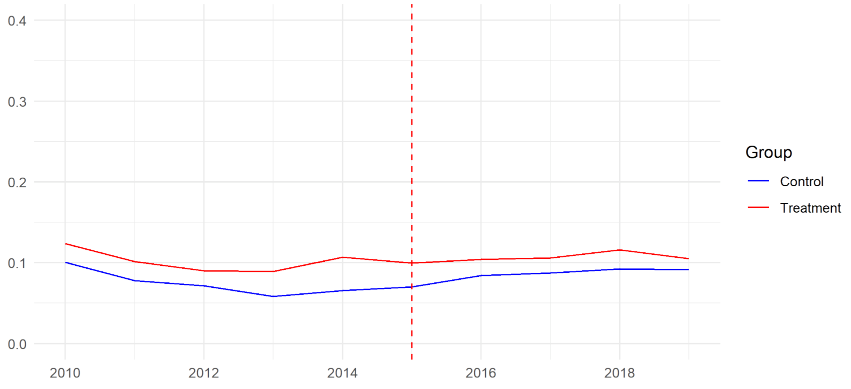


Figure 9.8: Parallel Trends - 12 years on Emotion (Angry) for 2015 as treatment

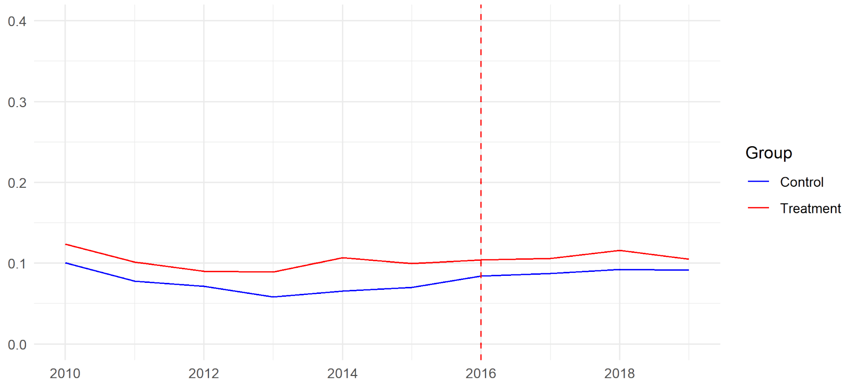


Figure 9.9: Parallel Trends - 12 years on Emotion (Angry) for 2016 as treatment

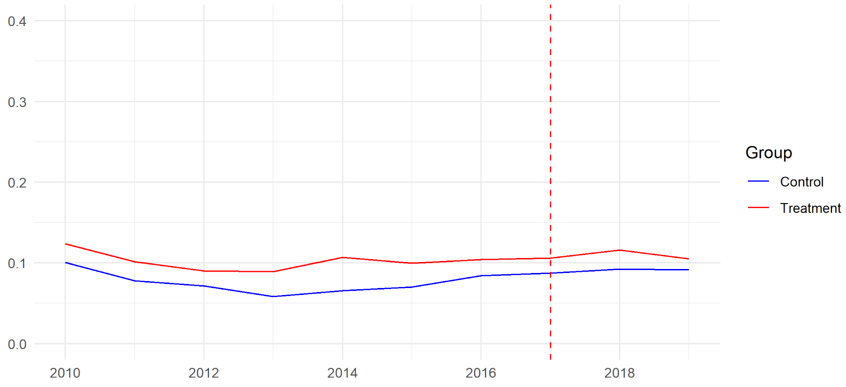
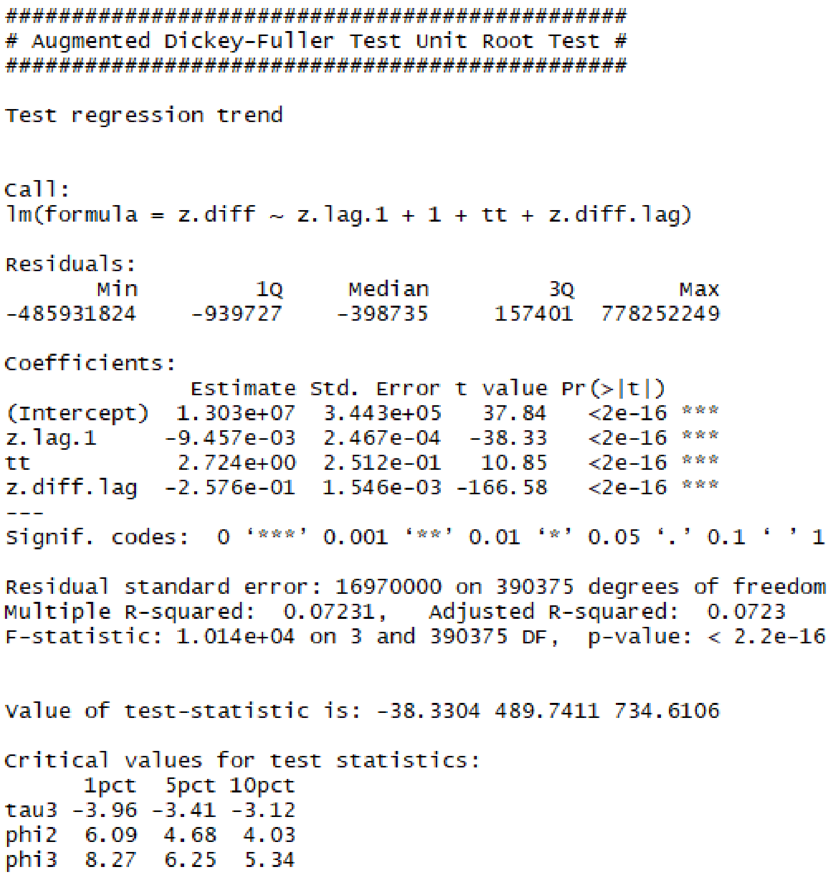


Figure 9.10: Parallel Trends - 12 years on Emotion (Angry) for 2017 as treatment



```
###############################################
# Augmented Dickey-Fuller Test Unit Root Test #
###############################################

Test regression trend


Call:
lm(formula = z.diff ~ z.lag.1 + 1 + tt + z.diff.lag)

Residuals:
       Min         1Q     Median         3Q        Max
-485931824    -939727    -398735     157401  778252249

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.303e+07  3.443e+05   37.84   <2e-16 ***
z.lag.1     -9.457e-03  2.467e-04  -38.33   <2e-16 ***
tt           2.724e+00  2.512e-01   10.85   <2e-16 ***
z.diff.lag  -2.576e-01  1.546e-03 -166.58   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 16970000 on 390375 degrees of freedom
Multiple R-squared:  0.07231,   Adjusted R-squared:  0.0723
F-statistic: 1.014e+04 on 3 and 390375 DF,  p-value: < 2.2e-16


Value of test-statistic is: -38.3304 489.7411 734.6106

Critical values for test statistics:
      1pct  5pct 10pct
tau3 -3.96 -3.41 -3.12
phi2  6.09  4.68  4.03
phi3  8.27  6.25  5.34
```

Figure 9.11: ADF Test