

ANALYZING AUDIENCE VIEWERSHIP OF OTT, TV, STREAMING PLATFORMS,  
AND SOCIAL MEDIA THROUGH COMPREHENSIVE INTELLIGENT –  
INTEGRATED PLATFORM

by

Chandramohan Puppala

DISSERTATION

Presented to the Swiss School of Business and Management Geneva

In Partial Fulfillment

Of the Requirements

For the Degree

DOCTOR OF BUSINESS ADMINISTRATION

SWISS SCHOOL OF BUSINESS AND MANAGEMENT GENEVA

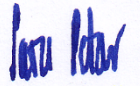
FEBRUARY, 2025

ANALYZING AUDIENCE VIEWERSHIP OF OTT, TV, STREAMING PLATFORMS,  
AND SOCIAL MEDIA THROUGH COMPREHENSIVE INTELLIGENT –  
INTEGRATED PLATFORM

by

Chandramohan Puppala

APPROVED BY



---

Prof.dr.sc. Saša Petar, Ph.D., Dissertation chair

RECEIVED/APPROVED BY:

---

Admissions Director

## **Dedication**

This thesis is dedicated to all those who challenged me during my career to bring out the best in me. To my parents, my wife and children, who instilled in me the values of perseverance and hard work, and to my media mentors, whose guidance and wisdom shaped my intellectual pursuits.

### **Acknowledgements**

This doctoral thesis is the result of years of dedication, emotional and psychological support from loved ones. I am deeply grateful to my mentor, Dr. Vijaykumar Varadarajan and Mr. Raghunandan Dhar for their unwavering guidance and encouragement. Special thanks to my friend Stephen Afrifa for his steadfast support and motivation. I also appreciate my media colleagues and BM batchmates for their insightful discussions and inspiration. This achievement would not have been possible without all those who contributed to my journey—thank you very much.

ABSTRACT  
ANALYZING AUDIENCE VIEWERSHIP OF OTT, TV, STREAMING PLATFORMS,  
AND SOCIAL MEDIA THROUGH COMPREHENSIVE INTELLIGENT –  
INTEGRATED PLATFORM

Chandramohan Puppala  
2025

Dissertation Chair: Dr. Vijayakumar Varadarajan

The rise of Over-the-Top (OTT) platforms, streaming services, television, and social media has transformed audience engagement, making sentiment analysis a crucial tool for understanding viewer opinions. Sentiment analysis, combined with machine learning techniques, enables the classification of audience sentiments into various sentiment categories. This study employs Random Forest (RF), Naïve Bayes (NB), and Support Vector Machine (SVM) to analyze sentiment trends in audience reviews. Data was collected from multiple sources, including social media discussions, streaming platform reviews, and TV audience feedback, resulting in a corpus of 2,000 text samples. The data was annotated, preprocessed, and classified using the three machine learning models. Random Forest outperformed the others, achieving 98.5% accuracy, 99.7% recall, and 99.8% precision, demonstrating its robustness in sentiment classification. Naïve Bayes followed with 95.7% accuracy, while SVM achieved 94.8% accuracy. Sentiment distribution analysis showed that positive sentiment dominated, followed by trust,

anticipation, and joy, while negative emotions such as fear, anger, and sadness were less frequent. A word cloud analysis further highlighted key themes related to content quality, storytelling, and viewer engagement. These findings suggest that Random Forest is the most effective model for sentiment classification in audience reviews. This study helps policy makers and stakeholders to analyze and make informed decisions on various cross-platform sentiment trends to enhance audience sentiment analysis across diverse media landscapes.

## TABLE OF CONTENTS

List of Tables.....	ix
List of Figures .....	x
CHAPTER I: INTRODUCTION .....	1
1.1 Background of the Study .....	1
1.2 Problem Statement .....	2
1.3 Purpose of Research .....	4
1.4 Research Questions .....	4
1.5 General Objective.....	5
1.6 Significance of the Study .....	5
1.7 Limitations of the Study.....	6
1.8 Organization of the Study .....	6
CHAPTER II: REVIEW OF LITERATURE.....	7
2.0 Introduction .....	7
2.1 Definition of Concepts .....	7
2.2 Streaming Services and the Consumption Rate .....	8
2.3 Audience Engagement and Behavioral Analytics .....	15
2.4 Application of Sentiment Analysis on Textual Data for Business Insights .....	22
2.5 Machine Learning Methods .....	30
2.6 Lexicon-based Approaches .....	37
2.7 Literature on Sentiment and Machine Learning Approaches.....	42
2.8 The Changing Media Landscape in India: The Need for an Integrated Platform.....	43
2.9 Related Works .....	49
2.10 Summary of Related Works .....	53
2.11 Problem to be Solved .....	54
CHAPTER III: METHODOLOGY .....	55
3.0 Introduction .....	55
3.1 Data Collection.....	55
3.2 The Conceptual Framework of the Study .....	57
3.3 Data Preprocessing .....	61
3.4 Exploratory and Qualitative Analysis .....	63
3.5 Artificial Intelligence Techniques .....	64
3.6 Performance Evaluation Metrics .....	67
3.7 Deployment of the Model.....	68
CHAPTER IV: EXPERIMENTAL RESULTS .....	70

4.0 Introduction .....	70
4.1 Outcome of the Sentiment Analysis Classification .....	70
4.2 Performance Analysis of the Machine Learning Classifiers .....	88
CHAPTER V: DISCUSSION .....	97
5.0 Introduction .....	97
5.1 Analysis of the Sentiment Distribution and Machine Learning Performance .....	97
5.2 Implications and Future Directions .....	102
CHAPTER VI: CONCLUSION, IMPLICATIONS, AND FUTURE RECOMMENDATIONS .....	108
6.1 Conclusion.....	108
6.2 Implications and Future Research .....	110
APPENDIX A SURVEY COVER LETTER.....	113
APPENDIX B INFORMED CONSENT .....	114
APPENDIX C INTERVIEW GUIDE.....	115
Interview Guide.....	115
Conclusion.....	117
REFERENCES .....	118

## LIST OF TABLES

<b>Table 2.6.1</b> Sentiment and machine learning studies on textual data.....	42
<b>Table 2.8.1</b> Summary of related works with the India landscape.....	48
<b>Table 2.10.1</b> Summary of related works.....	53
<b>Table 4.1.1.</b> Sample text and classification of the data. ....	76
<b>Table 4.2.1.</b> Performance of the classifiers on the data.....	90

## LIST OF FIGURES

<b>Figure 3.2.1.</b> The conceptual framework of the study.....	60
<b>Figure 3.7.1.</b> The Scrum phases for the implementation. ....	69
<b>Figure 4.1.1.</b> Word frequency of the data. ....	79
<b>Figure 4.1.2.</b> Word cloud of the dataset. ....	83
<b>Figure 4.1.3.</b> Sentiment score and effects from the dataset.....	88
<b>Figure 4.2.1.</b> Prediction outcome of the models.....	94
<b>Figure 4.2.2.</b> Optimal prediction outcome of the models.....	96

## CHAPTER I: INTRODUCTION

### 1.1 Background of the Study

The manner that media is consumed has changed as a result of global digitalization (Wang *et al.*, 2017). Better networks, more internet connections, new technological advancements, and the accessibility of smart gadgets have all contributed to the emergence of new on-the-top (OTT), streaming platforms, social media, which provides services to viewers directly via the internet (Atiqah *et al.*, 2021; Helm, 2021). An international network of linked computers and networks is known as the Internet (Ismond *et al.*, 2021). Broadband network bandwidth expansion has offered potential for OTT services to enter and undercut traditional broadcasting sectors. OTT services are those that are offered through online networks (Medina, Diego and Portilla, 2022). The name comes from the fact that these services are added on top of a customer's existing service. OTT does not necessitate the use of a broadcasting station, a cable connection, or a satellite television platform. The over-the-top (OTT) business is emerging as a fast-growing market, owing to the COVID 19 lockout. Though it is still in its early stages, its global market in 2019 was predicted to be US\$121.61 billion, and it is expected to reach US\$1039.03 billion in 2027, rising at a compound annual growth rate (CAGR) of 29.4% (He *et al.*, 2022).

The transition from traditional media to over-the-top (OTT) media, social media, and streaming platforms—especially with the advent of online gaming—has led to a competition among streaming service providers to draw and keep users. The global video streaming (SVoD) industry is anticipated to generate US\$95.88 billion in sales by 2023 (Hu *et al.*, 2022). This suggests that the market has a bright future. Traditional cable television subscriptions in the United States fell to 94 million households in 2018 (or 74%

of the projected 127 million US households) (Erkılıç and Erkılıç, 2022). This trend is particularly prevalent among the younger generation (18-34), who are far more inclined to use alternative video delivery platforms (Kim *et al.*, 2022). In Europe, national pay-TV penetration levels range from 24 to 97 percent (Scott and Paprocki, 2023), and losses have not reached the extreme levels seen in the United States; however, the majority of new revenue growth is coming from non-traditional Subscription Video on Demand (SVOD) providers (Kim, 2022). As traditional cable television providers continue to lose users, the future of OTT video providers appears bright. Recognizing a potential in the OTT video services industry, a range of entities, including established broadcasters, new content providers, and telecom firms, have introduced a variety of competitive products.

## **1.2 Problem Statement**

The expansion of the internet, along with the lockdown caused by the virus, has allowed OTT media to proliferate quickly (Feng, Wang and Chen, 2022). With the advent of online video streaming, various over-the-top (OTT) platforms are competing for the attention of viewers (Colombini and Duncan, 2023). Although services like as Netflix and Amazon Prime Video, and example of OTT service providers; have similar amounts of library content and user bases, they differ in terms of price structure, overall user experience, and catalogue availability (Harbin, 2023). The rise of Netflix in the United States exemplifies the proliferation of OTT video providers. It began as a DVD shipping service before launching its OTT video streaming service in 2007. According to data, there were over 800 OTT video providers worldwide as of late 2016 (Alcolea-Díaz, Marín-Lladó

and Cervi, 2022). The rise of these companies has jolted many countries' complicated vertically integrated broadcast television and cable businesses.

The rapid evolution of media consumption habits in the digital age has ushered in a new era for content delivery (Alkahtani and Aldhyani, 2021), through Over-The-Top (OTT) platforms, traditional TV, streaming services, and social media. As audiences have diversified their media consumption channels, understanding and measuring audience viewership across these platforms has become increasingly challenging and crucial for content creators, advertisers, and media organizations (Gascón-Vera and Marta-Lazo, 2023). This paradigm change has increased the complexity and segmentation of the viewing population, making complete media consumption data difficult to gather (Miller and Nelson, 2022). Even the number of Netflix subscribers, an OTT service provider, has overtaken the number of traditional cable TV subscribers in the United States (Huelin, 2022). Researchers have used the niche theory to compare not only new and old media (e.g., online/mobile news vs. traditional news), but also interpersonal media and services (e.g., cell phone vs. landline vs. instant messaging vs. email vs. text messaging), as well as competition patterns between two media.

Despite OTT, streaming platforms, and social media are continually expanding, the significance of conventional media cannot be underestimated. Therefore, a unified platform for accessing media consumption is essential for stakeholders to analyze and make informed decisions (Dabla, 2004). To analyze viewing on different channels, the study combines both qualitative and quantitative methodologies. To address this challenge, this study aims to develop a comprehensive integrated platform that leverages sentiment

analysis and machine learning techniques to measure audience viewership effectively. These would extract insightful meanings through the intelligent – powered platform for making informed decisions and business planning.

### **1.3 Purpose of Research**

The explosive rise of OTT, streaming platforms, and social media is having a massive impact on the entertainment business. However, cable networks and televisions are still used by a segment of the population. The latter's impact cannot be ignored, but it is empirical to investigate how these platforms' viewership is striving. Knowing the trends in all of these platforms can assist in making educated judgments and increasing productivity, and increase in subscription for multiple platforms to generate income.

### **1.4 Research Questions**

The number of OTT platform subscribers continues to outnumber those of cable networks and broadcast stations. As the population continues to use OTT and other streaming platforms, this raises the research questions as follows:

1. What framework can be developed to derive meaningful insights regarding audience viewership?
2. What model can be used to capture audience sentiments and reactions to content on different media platforms?
3. What intelligent machine models can be used to analyze the sentiment data and derive meaningful insights regarding audience viewership?
4. Which platform can be deployed to provide adaptable dashboard for real-time monitoring and visualization of audience viewership metrics?

5. How can the platform's effectiveness through case studies and comparison with existing measurement methods be validated?

## **1.5 General Objective**

The general objective of this study is to create a robust platform for measuring audience viewership across various media channels, including OTT, TV, streaming platforms, and social media.

### **1.5.1 Specific Objectives**

The specific objectives of this study are as follows:

1. To develop a framework to derive meaningful insights regarding audience viewership.
2. To develop a sentiment analysis model capable of capturing audience sentiments and reactions to content on different media platforms.
3. To implement machine learning algorithms to analyze the sentiment data and derive meaningful insights regarding audience viewership.
4. To deploy a user-friendly and intelligent-powered adaptable dashboard for real-time monitoring and visualization of audience viewership metrics.
5. To validate the platform's effectiveness through case studies and comparison with existing measurement methods.

## **1.6 Significance of the Study**

As the streaming industry becomes more crowded, OTT streaming providers must evaluate and enhance their engagement metrics more than before. These metrics provide

useful information about user behavior and preferences, helping streaming platforms to improve their content, user experience, and subscription growth.

### **1.7 Limitations of the Study**

The scarcity of data is a major impediment to the progress of this work. There is a scarcity of data in the literary area. Artificial intelligence (AI) techniques necessitate a massive amount of data to train these models. However, advanced or commercial data collecting software can be used to collect the data. A lack of financing is also a restriction; thus, this study aims to address it by utilizing questionnaire and social media data collection tools.

### **1.8 Organization of the Study**

The study is structured as follows: the first chapter is chapter 2, which discusses relevant works in the literature area. The second chapter is followed by the third, which is the methodology used for the study. In addition, the experimental results from the methods used are presented in Chapter 4. The "Discussion" section of Chapter 5 provides an in-depth review of the study. Chapter 6 brings the study to a close.

## CHAPTER II: REVIEW OF LITERATURE

### **2.0 Introduction**

This chapter examines prior research in the topic and identifies gaps. Furthermore, the concepts and terms in the domain are defined to provide the reader with an understanding of the study.

### **2.1 Definition of Concepts**

The concepts and definitions related with research are as follows.

- **Over-the-top:** OTT is an acronym that stands for "over-the-top" and refers to technology that streams material over the internet. Previously, a consumer would purchase a cable subscription, and their cable TV provider would be in charge of the supply and availability of programs.
- **Streaming Platforms:** Streaming platforms provide on-demand access to TV shows, movies, and other streaming material. Consider services such as Hulu, Netflix, Disney+, and Amazon Prime Video.
- **Television (TV):** TV is a type of mass media that involves the electrical transmission of moving pictures and sound from a source to a receiver. Television has had a significant impact on society by expanding the senses of vision and hearing beyond the constraints of physical distance.
- **Media:** The term "media" refers to all modes of communication, covering anything from printed paper to digital data. News, art, instructional content, and any type of

information that may reach or affect people, such as television, radio, books, magazines, and the internet, are all examples of media.

- **Social Media:** The phrase "social media" refers to websites and programs that emphasize communication, community-based input, engagement, content sharing, and collaboration.
- **Sentiment Analysis:** Sentiment analysis (also known as opinion mining) is a type of natural language processing (NLP) approach that determines whether input is positive, negative, or neutral. Sentiment analysis on textual data is frequently used to assist organizations in monitoring brand and product sentiment in consumer feedback and understanding customer demands.
- **Machine Learning:** Machine learning is a subfield of artificial intelligence (AI) and computer science that focuses on using data and algorithms to mimic how people learn, progressively improving its accuracy.
- **Artificial Intelligence:** Artificial intelligence (AI) is a broad field of computer science concerned with creating intelligent computers capable of doing activities that normally require human intelligence.

## **2.2 Streaming Services and the Consumption Rate**

Online video streaming has significantly altered the global media landscape and influenced watching habits all around the world. The worldwide video streaming industry has undergone remarkable growth in the previous decade, driven by significant jumps in internet usage, the broad availability of mobile devices, and the ever-increasing appeal of online video content. Over-the-top (OTT) video revenue is expected to reach 154 billion

US dollars in 2022, with the United States accounting for the greatest share of revenue globally. Given the fast expansion of international streaming services and the catalog of online video content, the number of OTT users globally is likely to reach new heights in the future. This incredible growth has seen the rise of platforms that not only challenge traditional television but redefine how media is consumed and produced.

The way we consume media has changed dramatically as a result of the rapid rate of technological advancement. Not long ago, the thought of never viewing TV in the sense of transmitted TV channels was unthinkable. Traditional forms of entertainment, such as cable television, were once the go-to medium for accessing television programs and movies. Nowadays, however, many people spend the majority, if not all, of their media consumption time on streaming services such as Netflix, Viaplay, Amazon Prime Video, Disney+, Hulu, and others (Iordache, Raats and Afilipoaie, 2022). These platforms have revolutionized how viewers interact with media by offering content on demand, free from the constraints of traditional broadcasting schedules. Streaming services now provide a diverse array of content, including movies, television shows, documentaries, sports, and user-generated content (UGC), catering to virtually every niche and interest. Consequently, viewers can personalize their entertainment experiences and curate content choices tailored to their tastes and preferences.

Streaming media is multimedia that has been sent, or "streamed," via the Internet for immediate consumption by a user. It is a common technique of transmitting and exchanging material, which has become central to the digital entertainment ecosystem. The term "streaming" was initially used in reference to computing in the 1970s, and as

previously stated, it gained popularity in the 1980s (Adway, 2023). At that time, the concept of streaming was primarily focused on local networks, where users could stream content within a single device or network environment. Streaming was relatively constrained in its scope, and technologies such as bandwidth limitations and data transfer rates posed significant barriers to broader implementation. However, over time, as broadband internet connections became more widespread, streaming began to evolve into a much more expansive medium capable of supporting global, on-demand content delivery.

In the early years, streaming had mostly a local connotation in its early stages. The focus was on watching media while recording on the same computer or sending data via a local network. At the time, wide area networks lacked the capacity and protocols to support anything other than basic text communication. As a result, content such as audio, video, and interactive media was limited in its reach, and users were reliant on physical storage devices for entertainment. Audio streaming technology advanced considerably during the 1990s (Fägersten and Bednarek, 2022), with innovations in audio compression algorithms and streaming protocols making it possible to deliver higher-quality content over dial-up internet connections. The radio sector was the first to be impacted by Internet streaming, as it was relatively easier to adapt pre-existing broadcast content to digital formats.

In the late 1990s and early 2000s, more and more established radio stations began to offer their regular broadcasts via the Internet, a popular service among diasporas who were looking for access to content from their home countries. These internet radio stations provided listeners with the ability to tune in from anywhere in the world, breaking down geographical barriers that previously limited access to traditional broadcast radio. The

music sector was the next to be impacted, with digital technologies enabling listeners to curate their own playlists and discover music based on personal preferences. Internet music radio stations began to evolve, offering users more control over what they listened to. These early services essentially transitioned from conventional radio stations to on-demand music streaming services.

Although Internet radio stations were permitted to operate through extensions of existing broadcast licensing agreements, early music streaming services faced a more ambiguous and complex legal landscape. The issue of copyright and licensing rights was a major challenge, and the recording industry initially resisted the idea of streaming music. Established record labels and artists feared that streaming would cannibalize CD sales and disrupt the traditional distribution model. Despite the opposition, digital music platforms began to gain traction, and services such as Pandora and Last.fm offered users access to personalized music streams. The real breakthrough came with the launch of Apple's legal download service, iTunes, in 2003 (Borchert and Seifert, 2023). iTunes was a game-changer, providing consumers with a legitimate, easy-to-use platform for purchasing and downloading music.

Even though the iTunes Store grew steadily throughout the 2000s, it was not enough to compensate for the decline in CD sales, and illegal downloading remained a significant concern for the music industry. Peer-to-peer file-sharing networks like Napster and LimeWire had paved the way for a culture of music piracy, which undercut the profitability of physical media. However, music streaming services such as MySpace and YouTube, which initially focused on user-generated content (UGC), also became important

players in the music field. YouTube, in particular, played a pivotal role in the rise of digital music consumption, allowing artists to reach global audiences and providing music fans with a platform for sharing and discovering new content.

The advent of Spotify in 2008 was a watershed point in the music streaming revolution. Spotify's freemium model—offering both free, ad-supported and premium, ad-free subscriptions—allowed users to stream an extensive catalog of music legally and without restrictions. Tidal, Apple Music, YouTube Music, and other platforms followed suit, offering their own subscription-based services to compete for market share. Streaming now accounts for more than 80% of music revenue in Scandinavian countries, and it became the primary source of music income globally by 2016 (Wayne and Castro, 2021). Streaming platforms have revolutionized the music industry, giving artists new ways to distribute and monetize their work while allowing consumers to access vast libraries of content on demand.

Despite the music industry's success with streaming, the video streaming industry has arguably had a more profound and far-reaching impact on global media consumption. Early projects like WebTV and interactive TV (iTV) offered glimpses of what was to come, but were ultimately unsuccessful due to technological limitations and an inability to meet consumer demands for convenience and content diversity (Vodičková, 2022). It was only with the development of broadband internet connections, better compression technologies, and streaming protocols that video streaming became a realistic and viable option in the mid-1990s. The rapid expansion of broadband networks around the world provided the necessary infrastructure for video streaming services to thrive. These advancements in

connectivity, along with the development of adaptive streaming technologies that adjusted video quality based on available bandwidth, made streaming video over the internet more reliable and scalable.

As in the music industry, the video streaming sector was initially dominated by download-only services and physical media like DVDs. The rise of DVD rental services like Blockbuster provided consumers with an alternative to traditional cable and satellite television services. However, technological advancements in video compression and adaptive streaming gradually solved many of the barriers to delivering high-quality video content over the internet. The adult film industry was among the first to embrace and develop new video streaming technologies, helping to drive innovation in the space (Maher and Cake, 2023). These innovations were crucial in making mainstream video streaming services feasible and more widely accessible.

The most significant milestone in video streaming occurred in 2005 with the introduction of YouTube. Created by three former PayPal employees, YouTube became an overnight sensation, offering users the ability to upload, share, and view video content for free. YouTube immediately became one of the most popular websites globally and continues to dominate the online video streaming space today. While YouTube initially gained its popularity by offering user-generated content, it has since evolved into a platform that includes professionally created content from film studios, television networks, and independent creators. The platform has served as a launching pad for countless viral trends, influencers, and content creators, further reshaping the entertainment landscape. Despite its massive growth and the increasing amount of high-quality content

on the platform, YouTube remains an outsider in the traditional film and television industries, with the content it hosts often being perceived as more informal or amateur compared to established media.

However, YouTube's impact cannot be overstated. The platform's influence in the world of media is so pervasive that it has fostered an entirely new era of digital entertainment. Content creators who might not have had access to traditional broadcasting channels can now build global followings and even monetize their content. Furthermore, YouTube's influence on other streaming platforms is significant, as its model of user-generated content has been adopted and expanded upon by competitors like TikTok and Vimeo, creating an ecosystem where users drive much of the content available to the public.

Video streaming also brought about a cultural shift in how people engage with media. With traditional television, viewers had to adhere to a set schedule for their favorite programs, often enduring commercial interruptions. Streaming services like Netflix, Hulu, and Amazon Prime Video have transformed television into a fully on-demand experience, where viewers are no longer bound by rigid programming schedules. This shift to on-demand content consumption has led to the rise of "binge-watching" culture, where viewers consume entire seasons of television series in one sitting. This form of media consumption has altered viewing patterns and is influencing how content is created, with many series now being structured for binge viewing rather than episodic releases.

In conclusion, the rapid rise of online video streaming services has fundamentally transformed the media consumption habits of global audiences. The shift from traditional

broadcast television and DVD rental services to on-demand streaming platforms has reshaped the entertainment industry. Services like Netflix, YouTube, and Spotify have not only changed how we access and consume media but have also introduced new business models, distribution strategies, and content creation opportunities. As internet infrastructure continues to improve and more people gain access to streaming services, the future of video and music streaming looks set to be even more dynamic, diverse, and influential in shaping the entertainment landscape. The impact of streaming technologies will continue to evolve and deepen, offering new possibilities for content creators and consumers alike.

### **2.3 Audience Engagement and Behavioral Analytics**

As digital platforms continue to dominate the entertainment landscape, understanding audience engagement and behavior has become increasingly crucial for platforms like OTT (Over-The-Top), TV, streaming services, and social media (Baccarne, Evens and Schuurman, 2013). These platforms are no longer just spaces for content delivery but have evolved into complex ecosystems that leverage data analytics, machine learning, and behavioral insights to enhance user experience, tailor content, and drive business growth (Z. Ren *et al.*, 2022). In this context, audience engagement and behavioral analytics play a pivotal role in predicting content consumption trends, influencing viewership, and optimizing platform strategies. This section will explore how data-driven techniques are used to study and predict audience behavior across these platforms and how such insights can improve the personalization of content, retention strategies, and the overall viewer experience.

### **2.3.1 Viewer Interaction Metrics**

Viewer interaction metrics are among the most important data points used to measure engagement on platforms such as OTT, TV, and social media (Tanrıöver, 2022). These metrics provide insight into how users interact with content and allow platforms to gauge the success of specific pieces of content. Common interaction metrics include likes, comments, shares, click-through rates (CTR), and watch time.

On social media platforms, likes and shares are often used as indicators of how emotionally resonant or compelling a piece of content is. On OTT platforms, watch time, including total minutes watched and binge-watching behaviors, is a key metric that signals user engagement and the effectiveness of content in keeping viewers invested (Kim *et al.*, 2022). In this digital age, platforms use real-time interaction data to dynamically adapt the user experience (Borchert and Seifert, 2023). For instance, if a show garners a significant number of social media mentions and interactions, platforms can quickly push that content to a wider audience.

Similarly, comments and reactions on social media platforms (such as Twitter or Facebook) provide a snapshot of the emotional tone of the content and how it resonates with users. Positive engagement typically correlates with a higher likelihood of viewers recommending or sharing content, thereby increasing its reach.

### **2.3.2 Personalization and Recommendation Systems**

Personalization is perhaps one of the most critical elements in ensuring user engagement and satisfaction on streaming platforms and social media (Ma and Sun, 2020). The advent of recommendation algorithms has revolutionized the way users discover content. These algorithms rely on collaborative filtering, content-based filtering, and hybrid models that leverage both user preferences and content characteristics to make predictions.

For example, platforms like Netflix, Hulu, and Spotify use machine learning models to personalize content recommendations by analyzing past viewing history, user ratings, and search behaviors. These systems can predict what a user is most likely to watch next, improving content discoverability and reducing churn rates. The effectiveness of recommendation systems is crucial in OTT and streaming platforms, where an overwhelming number of shows and movies are available, making it difficult for users to find relevant content without assistance (He *et al.*, 2022).

Furthermore, sentiment analysis plays a vital role in recommendation systems. By analyzing the sentiment of user reviews, social media discussions, and comments, platforms can fine-tune their recommendation systems to prioritize content that aligns with the emotional preferences of users. For example, a user who consistently watches romantic comedies may be more likely to enjoy a newly released romantic film with positive sentiment or critical acclaim.

### **2.3.3 Content Consumption Patterns**

Content consumption patterns refer to how users engage with content, including viewing frequency, preferred genres, and binge-watching behavior. These patterns can vary widely across platforms. On traditional TV, viewers often follow a weekly episodic schedule, while OTT platforms like Netflix and Amazon Prime Video have capitalized on binge-watching behavior by releasing entire seasons of shows at once (Pedrero-Esteban, Terol-Bolinches and Arense-Gómez, 2023).

Binge-watching, defined as watching multiple episodes of a show in a single sitting, has become a defining characteristic of OTT platforms. Streaming platforms encourage this behavior by utilizing features like auto-play and suggesting related content (Wayne, 2022). However, binge-watching has implications for content retention, as users are more likely to become invested in a series if they can watch it continuously without interruption. Social media platforms also influence content consumption by providing instant feedback on popular trends. A viral video or meme on a platform like TikTok or Twitter can cause a ripple effect that drives users to streaming platforms to watch a specific show or film. This creates a synergistic relationship between traditional TV, OTT, and social media, where trends on one platform can influence viewership on another.

### **2.3.4 Impact of Social Media on OTT and TV Viewership**

Social media has become a significant driver of audience engagement across traditional TV and OTT platforms. Platforms like Twitter and Instagram

allow viewers to share their thoughts, reactions, and fan theories about TV shows, movies, and live broadcasts, influencing their social circles and contributing to a show's viral success (Alcolea-Díaz, Marín-Lladó and Cervi, 2022). In fact, social TV is a term used to describe the intersection of television and social media, where users engage with others around live programming (Feng, Wang and Chen, 2022). For instance, live-tweeting during a series premiere or finale has become a cultural phenomenon, particularly for reality TV shows or event programming like the Super Bowl. Similarly, OTT platforms use social media to amplify the visibility of new content. The relationship between social media trends and viewership is increasingly reciprocal; as more users discuss a show on platforms like Twitter or Facebook, the more likely it is that others will tune in to see what the buzz is about. Social media has also reshaped content promotion. Influencers and celebrities with large followings can significantly impact the popularity of TV shows and movies by sharing their opinions or endorsing content. Streaming platforms like Netflix and YouTube work closely with influencers to promote original content, leveraging their reach to attract a more targeted audience (Wayne and Castro, 2021).

### **2.3.5 Behavioral Segmentation and Audience Profiling**

One of the core applications of audience engagement and behavioral analytics is audience profiling, which is used to categorize users into distinct behavioral segments (Medina, Diego and Portilla, 2022). This segmentation allows platforms to provide more personalized content recommendations, targeted advertising, and content offerings tailored to specific audience groups. Streaming

platforms often employ segmentation techniques to understand users based on their content preferences, viewing patterns, and demographic factors (Chan-Olmsted, 2019). For instance, users who watch a lot of action films or thrillers may be categorized as part of an "action movie lover" segment. These insights help platforms recommend new releases within those genres. Behavioral segmentation is also useful for targeted marketing, as platforms can serve ads or suggest subscriptions based on the user's viewing history and interests.

### **2.3.6 Psychographics and Consumer Psychology**

Psychographics refers to the study of consumer psychology, values, attitudes, and emotions that influence decision-making. In the context of OTT platforms and social media, psychographics provides insight into why certain types of content resonate with specific audiences (Kovačević and Perišin, 2022). For example, content that triggers emotional contagion (e.g., inspiring or motivational stories) can generate a positive response among viewers, increasing their engagement and likelihood of sharing content (Killian and McManus, 2015). On the other hand, content that evokes strong negative emotions (such as fear or anger) may go viral due to its emotional impact, but this may not always translate into long-term viewer loyalty. Understanding psychographics is important for streaming platforms to develop content that aligns with the emotional needs of different viewer segments. Whether it's creating a heartwarming drama or a thrilling mystery, knowing the emotional triggers of an audience can help platforms increase engagement and retention.

### **2.3.7 Audience Retention Strategies**

For OTT and streaming platforms, audience retention is critical to business success. While acquiring new subscribers is important, retaining existing users is equally essential (Nauta *et al.*, 2022). Audience retention strategies include exclusive content, loyalty programs, and subscription model optimization. Exclusive content, such as original series and films, has become a key differentiator for platforms like Netflix, Disney+, and Amazon Prime. By offering content that cannot be found elsewhere, these platforms encourage users to stay subscribed for the long term. Similarly, platforms use loyalty programs and tiered subscription models to retain viewers and incentivize longer subscriptions.

Audience engagement and behavioral analytics are fundamental components of understanding and optimizing user behavior on OTT, TV, streaming platforms, and social media (Iordache, Raats and Afilipoaie, 2022). By analyzing user interactions, personalizing recommendations, studying content consumption patterns, and leveraging social media trends, platforms can create a more engaging and satisfying experience for their audiences (Batik and Demir, 2022). As the media landscape continues to evolve, these data-driven strategies will remain critical to shaping the future of content consumption and delivery. Future developments in machine learning and deep learning will likely provide even more sophisticated tools for understanding audience behavior, leading to more personalized, effective, and engaging platforms.

## **2.4 Application of Sentiment Analysis on Textual Data for Business Insights**

Real-time sentiment analysis has become an indispensable technique in the realm of data-driven decision-making, especially for businesses that wish to gain insights into public opinion and consumer sentiment in real-time (Schwenk, Wyss and Aubry, 2025). The technique employs machine learning (ML) algorithms to identify and analyze subjective opinions expressed across various social media platforms, internet forums, customer feedback, and other textual data sources. Real-time sentiment analysis plays a crucial role in understanding public perception, enabling organizations to respond quickly to consumer concerns, monitor brand reputation, and track market trends (Hossain *et al.*, 2025). This process involves extracting sentiment from text, categorizing it as positive, negative, or neutral, and using this information to gain actionable insights for business strategies and customer engagement.

For businesses, the value of real-time sentiment analysis extends beyond simple monitoring. It provides actionable insights that can guide marketing campaigns, inform product development, and improve customer service (Xiang *et al.*, 2023). For example, companies can track mentions of their brand, products, or services to gauge customer satisfaction and loyalty. Furthermore, sentiment analysis can identify emerging trends, highlight potential public relations issues, and help businesses navigate competitive landscapes (Suganthi, 2024). This is particularly important in today's digital economy, where customer opinions are freely expressed in real-time on platforms like social media, product review sites, and blogs.

In practical terms, real-time sentiment analysis is used to monitor brand mentions for market intelligence, track specific keyword mentions for research purposes, and even dissuade harmful activities like cyberbullying (Shiva *et al.*, 2021). The ability to monitor real-time sentiment provides businesses with an opportunity to take swift action in addressing negative feedback or capitalizing on positive sentiment. This can include adjusting marketing strategies, improving customer service, or making product enhancements based on customer feedback.

Moreover, real-time sentiment analysis allows businesses to gain insights into public sentiment not only for their own brands but also for their competitors. By monitoring online discussions about rival brands, companies can better understand market positioning, identify opportunities for differentiation, and detect potential gaps in the market (Chapola *et al.*, 2023). In competitive industries, where customer perceptions and preferences change rapidly, real-time sentiment analysis can serve as a critical tool for staying ahead of market trends.

The application of sentiment analysis is widespread across various domains, including customer service, marketing, and public relations (Bonta, Kumaresh and Janardhan, 2019). For instance, many companies leverage sentiment analysis tools to monitor customer interactions through chatbots or customer service interactions, ensuring that they are providing timely, relevant, and positive support. Additionally, sentiment analysis is used extensively in marketing and advertising to measure the effectiveness of campaigns and determine how advertisements are being received by different audience segments (Rivas *et al.*, 2022). It also serves as a valuable tool for social media managers

who must assess public sentiment regarding viral trends, hashtags, and current events, allowing them to respond quickly and effectively to shifts in public opinion.

One of the primary challenges in real-time sentiment analysis is accurately determining the polarity of text—whether it is positive, negative, or neutral. This seems straightforward in theory, but in practice, it can be complicated by several factors such as sarcasm, irony, or ambiguity in language (Alqurashi, 2022). For example, a sentence like “I just love waiting in line for hours at this store” may appear positive at first glance but is, in fact, a negative statement. This kind of nuance requires advanced machine learning techniques and natural language processing (NLP) to decipher accurately.

To better understand sentiment analysis, it is helpful to break the process down into three primary levels of analysis: sentence-based, document-based, and aspect-based (Faruque *et al.*, 2021). These three layers allow for varying degrees of sentiment extraction, from analyzing individual sentences to examining broader documents or specific aspects of products or services. Each of these layers offers a different perspective on sentiment analysis, enabling organizations to gain a comprehensive understanding of consumer sentiment.

#### **2.4.1 Sentence-Based Sentiment Analysis**

The sentence-based approach is the most basic form of sentiment analysis and is often used in applications where the goal is to determine the sentiment of individual sentences within a larger body of text (Ponce, Cruz and Andrade-arenas, 2022). This method involves analyzing each sentence separately to classify it as positive, negative, or neutral. Sentence-based sentiment analysis is useful for

identifying specific customer opinions on individual topics or statements (Model, Ou-yang and Chou, 2022). For example, in a customer review, one sentence might express satisfaction with a product, while another might express dissatisfaction with customer service. By analyzing each sentence independently, the sentiment of each aspect of the review can be understood in detail.

Sentence-based analysis is particularly useful in situations where precise feedback is required on specific features or attributes of a product or service (Maharani and Effendy, 2022). In the context of real-time sentiment analysis, this approach allows businesses to quickly identify both positive and negative reactions and address specific issues in a timely manner. However, the limitation of sentence-based analysis lies in its inability to capture the overall tone or context of the text, as it treats each sentence in isolation. This means that competing ideas or conflicting opinions within a document may not be fully captured, which can lead to inaccurate sentiment interpretation.

#### **2.4.2 Document-based Sentiment Analysis**

In contrast to sentence-based analysis, document-based sentiment analysis looks at the overall sentiment of an entire document or text. This approach aims to determine the general sentiment conveyed in the document, considering all the sentences collectively (Garg and Sharma, 2022). Document-based analysis is valuable for understanding the overarching sentiment in longer texts, such as customer reviews, blog posts, or social media threads. By aggregating the sentiment of individual sentences, document-based sentiment analysis provides a broader

view of how a brand, product, or topic is perceived overall (Catelli, Pelosi and Esposito, 2022).

This approach is particularly useful for gaining insights from large volumes of text, such as monitoring brand sentiment over time. For example, a company may track customer sentiment over several months to determine whether customer perceptions are improving or declining. Document-based sentiment analysis can also help identify the emotional tone of an article or social media post, allowing businesses to adjust their messaging or take corrective actions if necessary. However, document-based sentiment analysis may still miss nuances or competing viewpoints within the text, as it focuses on the overall sentiment rather than addressing specific aspects of the content.

#### **2.4.3 Aspect-based Sentiment Analysis**

Aspect-based sentiment analysis (ABSA) is the most advanced form of sentiment analysis and involves examining specific aspects or features of a product, service, or brand. This approach is particularly useful when a business needs to evaluate customer sentiment toward particular elements of its offerings, such as the quality of customer service, product performance, or price (Uddin and Hafiz, 2022). By breaking down sentiment into distinct aspects, businesses can gain a deeper understanding of which areas are driving positive or negative opinions.

For example, in an online review of a smartphone, customers might provide feedback on various aspects such as battery life, screen quality, and camera performance. Aspect-based sentiment analysis enables the identification of

sentiment associated with each of these aspects, helping the company understand what customers like or dislike about specific features. This type of analysis is invaluable for product development, as it allows businesses to prioritize improvements based on customer feedback about specific aspects (Al-Hashedi *et al.*, 2022).

One of the challenges of aspect-based sentiment analysis is that it requires the ability to accurately identify and categorize aspects within text. For example, in a review of a hotel, the aspects could include room cleanliness, location, service quality, and amenities. The challenge lies in ensuring that the sentiment expressed toward each aspect is accurately captured, as different aspects may evoke different emotions or opinions (Jin, 2020). Moreover, aspect-based sentiment analysis often requires more advanced NLP techniques to recognize and extract the relevant aspects of a product or service.

#### **2.4.4 Practical Applications of Real-time Sentiment Analysis**

For research in the domain of business administration (DBA), real-time sentiment analysis offers numerous practical applications. As businesses strive to gain a competitive edge in the marketplace, understanding consumer sentiment becomes a key strategic tool. Sentiment analysis can provide valuable insights into customer preferences, brand perception, and market trends, which are essential for data-driven decision-making (Rintyarna *et al.*, 2022).

#### **2.4.5 Brand Monitoring and Reputation Management**

One of the most common uses of sentiment analysis is brand monitoring, where companies track online mentions of their brand, products, or services in real-time. Sentiment analysis tools can identify whether these mentions are positive, negative, or neutral, allowing businesses to address customer concerns quickly (Nguyen, Al and Academy, 2018). For example, if a company receives negative feedback on social media, sentiment analysis can help the company identify the issue and respond in real-time, potentially turning a dissatisfied customer into a loyal one. In the context of DBA research, exploring how companies use sentiment analysis to manage their reputation and improve customer satisfaction can yield insights into best practices for reputation management in the digital age.

#### **2.4.6 Customer Feedback and Product Improvement**

Another practical application is in analyzing customer feedback for product improvement. By analyzing sentiment at the aspect level, businesses can identify areas where their products or services are performing well and areas where they need to improve (Dashtipour *et al.*, 2016). For example, sentiment analysis could reveal that customers are generally happy with a product's design but dissatisfied with its battery life. This information can be used to guide product development and improve customer satisfaction. DBA research could explore how companies integrate sentiment analysis into their product development cycles to make data-driven decisions about product features and enhancements.

#### **2.4.7 Competitive Analysis**

Real-time sentiment analysis also offers significant value in competitive analysis. By monitoring sentiment around competitors' brands, products, or services, businesses can identify market trends, gaps, and opportunities for differentiation (Zhang, Gan and Jiang, 2014). For example, if competitors are receiving negative feedback about a particular aspect of their service, a business could leverage this insight to position its own offerings more favorably. DBA research can explore how sentiment analysis tools are used by companies to gain competitive intelligence and make strategic decisions in dynamic market environments.

#### **2.4.8 Social Media Engagement and Marketing**

Social media platforms are a rich source of real-time sentiment data. Businesses use sentiment analysis to gauge public opinion on marketing campaigns, track the effectiveness of advertisements, and measure brand sentiment (Hoang, Bihorac and Rouces, 2019). By analyzing sentiment surrounding a hashtag or campaign, businesses can adjust their messaging or strategies to better resonate with their target audience. DBA research can investigate how companies use sentiment analysis to optimize marketing strategies and enhance customer engagement in the digital age.

Real-time sentiment analysis is a powerful tool that enables businesses to understand public opinion and consumer sentiment in real-time. By leveraging machine learning and natural language processing, businesses can gain actionable

insights into how they are perceived by customers, identify areas for improvement, and adjust their strategies accordingly (Kusal *et al.*, 2021). The three primary layers of sentiment analysis—sentence-based, document-based, and aspect-based—offer different levels of insight, each serving a specific purpose in understanding sentiment. In the context of DBA research, real-time sentiment analysis offers significant practical value in areas such as brand monitoring, product improvement, competitive analysis, and social media marketing. As businesses continue to embrace sentiment analysis tools, the ability to analyze and respond to consumer sentiment in real-time will play an increasingly important role in shaping business strategies and driving competitive advantage in today's fast-paced digital economy.

## **2.5 Machine Learning Methods**

Text classification is a vital component of machine learning, and supervised algorithms play a key role in categorizing content into various predefined classes such as positive, negative, or neutral (Singh and Glińska-Neweś, 2022). These methods rely on training a model using labeled data, where each data point is associated with a class label. The algorithm identifies patterns and characteristics that distinguish different classes and learns to generalize from these patterns to classify previously unseen data. The ability to classify text has wide applications across various domains, including sentiment analysis, spam detection, topic classification, and more.

Supervised text classification typically involves the extraction of features from the raw text, which can include keywords, frequency of specific terms, or syntactic structures. These features serve as inputs to the classification algorithm, which then infers a function

capable of categorizing new, unseen examples. Common machine learning methods for text classification include support vector machines (SVM), decision trees, k-nearest neighbors (KNN), and naive Bayes classifiers. Despite their usefulness, these models can struggle with complex language structures, especially in large datasets, where traditional approaches may be limited by their inability to capture intricate relationships within the text.

A more recent and increasingly popular model for text classification tasks is the Convolutional Neural Network (CNN) (Maharani and Effendy, 2022). CNNs are a type of deep learning model that has gained significant attention due to their remarkable success in both image recognition and natural language processing (NLP) tasks. Initially, CNNs were designed for processing grid-like data, such as images, where they excelled at learning hierarchical spatial features. However, their application has since expanded to NLP tasks, where CNNs have been shown to effectively handle sequential data like text. This expansion is due to their ability to capture local patterns in data and apply them to more complex structures.

In the context of NLP, CNNs operate by using convolutional layers, which allow the model to capture local dependencies between words in a sequence of text (Ray and Chakrabarti, 2022). This is particularly useful for tasks such as sentiment analysis or document classification, where understanding local word patterns is essential. The convolutional layer in CNNs performs operations that apply a set of filters across the input text, detecting features like word combinations or n-grams that are useful for determining the sentiment or category of the text.

### 2.5.1 CNNs in Text Classification

CNNs have recently been recognized for their ability to successfully process natural language data (Ray and Chakrabarti, 2022). While traditional text classification models like SVM or Naive Bayes are based on a bag-of-words model, which treats words as independent entities, CNNs are capable of capturing more complex relationships between words by considering their spatial arrangement within sentences or documents. This makes CNNs particularly well-suited for tasks where context is important for classification, such as determining the sentiment of a review or identifying the topic of an article (Shang *et al.*, 2023).

In CNN-based models for text classification, the first step typically involves representing words as vectors (Xu *et al.*, 2022). These vectors are often pre-trained word embeddings like Word2Vec, GloVe, or FastText, which capture semantic meaning and relationships between words based on their co-occurrence patterns in large corpora. Once the text is transformed into vector representations, the CNN model can begin processing it through multiple layers. A convolutional layer in a CNN for text classification uses a sliding window mechanism, applying a set of filters (also known as kernels) over the input sequence (Lin *et al.*, 2023). Each filter is designed to detect specific patterns in the text, such as word pairs, phrases, or specific combinations of words that are indicative of a certain class. These filters scan the text in a sliding window fashion, performing element-wise multiplication followed by a summation operation to generate feature maps. The resulting feature maps highlight the presence of certain patterns or features in the text.

The pooling layer that follows the convolutional layer plays a critical role in reducing the dimensionality of the output and allowing the model to generalize better (Lengkeek, van der Knaap and Frasinca, 2023). Pooling layers are typically designed to take the maximum or average value from a group of neighboring values in the feature map, thereby reducing the complexity of the model while retaining important features. This process also helps to make the model more robust to variations in word order and position, which is important for text classification tasks where the meaning can remain unchanged despite changes in word order (Sherif *et al.*, 2023).

### **2.5.2 Deep Learning Architecture and Its Impact on NLP**

CNNs, as part of a broader class of deep learning models, have shown tremendous promise in various NLP applications (Firoozabadi *et al.*, 2020). Deep learning models, including CNNs, utilize hierarchical structures with multiple layers of nonlinear transformations to learn complex representations of the input data (Ham *et al.*, 2022). The deep architecture of CNNs enables them to extract high-level abstract features from raw input data, progressively building more complex representations as the data passes through successive layers. In traditional machine learning models, feature extraction is a manual process, requiring domain expertise to define and select the most relevant features (Li *et al.*, 2023). In contrast, deep learning models like CNNs automate feature extraction, learning the most important features from data in an unsupervised or semi-supervised manner. This ability to learn from raw data with minimal manual intervention is one of the

reasons deep learning has seen such widespread success in a range of applications, from image recognition to speech processing and NLP.

The combination of CNNs and deep learning has enabled significant advancements in NLP, especially in tasks that require capturing long-range dependencies and complex patterns in text (Kumar, Kumar and Soman, 2019). For example, in sentiment analysis, CNNs can capture patterns in word combinations, such as "not good" or "very happy," that are important for determining the overall sentiment of a text. Similarly, in document classification tasks, CNNs can learn to identify specific topics by recognizing key phrases or terms that indicate a certain category.

Another important aspect of CNNs in text classification is their ability to process variable-length input sequences (Alessandrini *et al.*, 2023). In NLP, text can vary widely in length, from a short tweet to a lengthy news article. Traditional machine learning models often require fixed-length inputs, which can lead to information loss when handling variable-length text. CNNs, however, can operate on input sequences of varying lengths, processing each sequence through multiple layers of convolution and pooling to extract features at different levels of abstraction.

### **2.5.3 Advantages of CNNs in Text Classification**

The advantages of using CNNs for text classification are numerous. One of the main benefits is their ability to learn hierarchical feature representations (Gajjar *et al.*, 2024). In traditional machine learning models, features are typically

predefined and do not evolve over time. CNNs, on the other hand, learn to extract features directly from the input data, which allows them to adapt to changing patterns and relationships in text. This is particularly useful in dynamic fields such as social media monitoring, where trends and language usage can shift rapidly. Another advantage of CNNs is their computational efficiency (Rezaul *et al.*, 2024). CNNs can be parallelized effectively, making them suitable for large-scale text classification tasks. As a result, CNNs are capable of processing massive amounts of textual data in relatively short periods of time, a crucial factor for applications such as real-time sentiment analysis or social media monitoring, where speed is essential.

Moreover, CNNs are relatively robust to noise in the input data. For example, in sentiment analysis, a CNN can still classify a text correctly even if there are spelling mistakes, informal language, or other types of noise present (Olivieri *et al.*, 2024). This is because CNNs focus on local features rather than relying on exact word matches, which makes them more tolerant of variations in the input.

#### **2.5.4 Challenges and Considerations**

While CNNs offer many advantages for text classification, they also present certain challenges. One of the main challenges is the need for large labeled datasets for training. CNNs, like most deep learning models, require substantial amounts of data to learn meaningful representations and generalize well (Wang *et al.*, 2017). In domains where labeled data is scarce or expensive to obtain, training a CNN model can be difficult and resource-intensive. Furthermore, CNNs are not always

the best choice for tasks that require modeling long-term dependencies in text, such as language translation or document summarization (Taboada, Brooke and Voll, 2022). For these tasks, recurrent neural networks (RNNs) or transformers may be more appropriate, as they are better suited for capturing sequential dependencies over long distances.

Another challenge with CNNs in text classification is interpretability. Deep learning models, including CNNs, are often considered "black boxes" because it can be difficult to understand why a model makes a particular prediction (Kawade and Oza, 2017). While techniques such as saliency maps or attention mechanisms can help provide some insight into which parts of the text are influencing the model's decisions, CNNs still lack the transparency of simpler models like decision trees or SVMs, which can be a drawback in applications that require explainability.

Text classification using machine learning, particularly through Convolutional Neural Networks (CNNs), has become a powerful approach for analyzing and categorizing textual data in a variety of domains, including sentiment analysis, topic classification, and spam detection (Jurek, Mulvenna and Bi, 2015). CNNs offer several advantages over traditional machine learning methods, including the ability to automatically learn hierarchical features, process variable-length input sequences, and handle large datasets efficiently. Their success in natural language processing is largely due to their ability to capture local dependencies in text and their adaptability to changing patterns (Lye and Teh, 2021). However, despite these advantages, there are challenges associated with

CNNs, such as the need for large datasets and the difficulty in interpreting the model's decision-making process. As with any machine learning model, careful consideration of the specific task and available resources is necessary when choosing CNNs for text classification tasks (Y. Ren *et al.*, 2022). Nonetheless, the remarkable success of CNNs in various NLP applications underscores their potential as a transformative tool in the field of text analysis.

## **2.6 Lexicon-based Approaches**

Polarity refers to the emotional tone or sentiment expressed by a word, which can evoke positive, negative, or neutral reactions in the mind of the reader (Xu *et al.*, 2022). Understanding the polarity of words is a fundamental aspect of sentiment analysis, which seeks to determine the sentiment or opinion conveyed in a piece of text. Sentiment analysis is widely used in various applications, such as social media monitoring, brand reputation management, and customer feedback analysis, to gain insights into public perception.

In sentiment analysis, polarity plays a central role in classifying and understanding the emotional tone of a text (Shang *et al.*, 2023). Words are often categorized as positive, negative, or neutral, depending on the sentiment they convey. This classification allows for a clearer understanding of the emotions embedded in the text and can be used to assess overall sentiment. For example, positive words such as "happy" or "excellent" might indicate a favorable sentiment, while negative words like "angry" or "poor" suggest dissatisfaction or disapproval.

### **2.6.1 Approaches in Lexicon-Based Sentiment Analysis**

A common approach to sentiment analysis is lexicon-based sentiment analysis, which relies on predefined dictionaries or lexicons that contain words associated with specific sentiment polarities (H. Manguri, N. Ramadhan and R. Mohammed Amin, 2020). These lexicons classify words based on their emotional tone, making it possible to analyze the sentiment of a given text by looking up individual words in the lexicon. Lexicon-based sentiment analysis is simple to implement and can be effective in many contexts, especially when dealing with structured text or when the goal is to quickly gauge the overall sentiment of a document (Guerini, Gatti and Turchi, 2013).

There are two primary types of lexicons used in sentiment analysis: weighted and unweighted. In a weighted lexicon, words are assigned specific scores or weights, indicating the intensity of their sentiment (Singh, Sawhney and Kahlon, 2018). For example, the word "great" may be assigned a higher positive weight than the word "good," reflecting the stronger positive sentiment conveyed by "great." On the other hand, an unweighted lexicon simply categorizes words as positive, negative, or neutral, without assigning a score based on their intensity. Both approaches have their advantages and can be chosen based on the specific needs of the sentiment analysis task.

### **2.6.2 Polarity in Lexicon-Based Sentiment Analysis**

The process of polarity detection in lexicon-based sentiment analysis involves identifying and categorizing words in a given text according to their

sentiment. When analyzing a sentence, the system examines each word to determine its polarity based on the sentiment lexicon. For instance, consider the sentence, "Good people sometimes have bad days." In this example:

- *The word "Good" would be classified as positive.*
- *The word "Bad" would be classified as negative.*

The words "people," "sometimes," "have," and "days" would be classified as neutral because they do not convey a clear sentiment.

Once the words are categorized, the overall sentiment of the sentence can be determined by counting the number of positive and negative words and calculating a sentiment score. The sentiment score can be computed by subtracting the number of negative words from the number of positive words. If the result is positive, the sentence expresses a positive sentiment; if the result is negative, the sentence expresses a negative sentiment; and if the result is zero, the sentiment is neutral.

In this example, the sentence "Good people sometimes have bad days" contains one positive word ("Good") and one negative word ("Bad"). If we assign each word a value of +1 for positive and -1 for negative, the overall sentiment score would be:

- *Positive score = +1 (for "Good")*
- *Negative score = -1 (for "Bad")*

The overall sentiment score would be 0, suggesting a neutral sentiment. This approach allows sentiment analysis to be easily automated and applied to large volumes of text, such as customer reviews or social media posts.

### **2.6.3 Challenges in Lexicon-Based Sentiment Analysis**

While lexicon-based sentiment analysis is a useful technique, it does have some limitations. One of the key challenges is handling negations and context. In the example above, the phrase "Good people sometimes have bad days" is relatively straightforward, but in other cases, the sentiment of a word can be reversed or modified by surrounding words. For instance, the phrase "I don't like this product" contains the word "like," which would typically be classified as a positive word. However, the presence of the negation "don't" changes the sentiment of the sentence to negative. Similarly, the phrase "I really enjoy this movie" uses the word "enjoy," which is positive, but the intensifier "really" amplifies the sentiment.

Lexicon-based methods can also struggle with polysemy, where a word has multiple meanings depending on the context (Balakrishnan Gokulakrishnan, Pavalanathan Priyanthan, Thiruchittampalam Ragavan, Nadarajah Prasath, 2012). For example, the word "bitter" can refer to a negative taste or emotion, but in the context of "a bitter victory," it could refer to a sense of sadness or regret despite success. These complexities can make it difficult to accurately classify sentiment based solely on individual words. Moreover, lexicon-based sentiment analysis may fail to capture the nuances of language, such as sarcasm or irony. In sarcastic statements, the words may appear to be positive, but the intended sentiment is

negative. For example, "Oh, great, another flat tire" contains the word "great," which would typically be classified as positive. However, the context of the sentence suggests that the speaker is actually frustrated, and the sentiment is negative. This limitation highlights the need for more advanced techniques that can account for context and sentence structure.

#### **2.6.4 Enhancing Lexicon-Based Sentiment Analysis**

To address some of the challenges associated with lexicon-based sentiment analysis, researchers have developed several techniques to improve its accuracy. One such method is the use of sentiment lexicons that include domain-specific words or context-sensitive features. For example, a lexicon for movie reviews may include words such as "entertaining" or "boring," while a lexicon for product reviews may include terms like "durable" or "cheap." These domain-specific lexicons can help improve the accuracy of sentiment classification by considering the specific context of the text.

Another approach is to combine lexicon-based sentiment analysis with other techniques, such as machine learning models. For example, a machine learning classifier can be trained on labeled data to recognize the sentiment of a text based on its overall structure and context, while the lexicon can be used to identify individual words with strong sentiment. This hybrid approach allows for more accurate sentiment classification by leveraging the strengths of both methods. Additionally, context-aware sentiment analysis techniques, such as using dependency parsing or word embeddings, can help capture the relationships

between words and improve the handling of negations, intensifiers, and other contextual factors. For instance, word embeddings like Word2Vec or GloVe represent words in a continuous vector space, where words with similar meanings are placed close together. These embeddings can help capture the meaning of words in context, providing a more nuanced understanding of sentiment.

Lexicon-based sentiment analysis is a valuable tool for understanding the emotional tone of text by classifying words as positive, negative, or neutral. By leveraging sentiment lexicons, this approach can quickly provide insights into public opinion, customer feedback, or brand perception. However, challenges such as handling negations, context, and polysemy can limit the effectiveness of lexicon-based methods.

To improve sentiment analysis, it is essential to combine lexicon-based techniques with more advanced methods, such as machine learning models and context-aware techniques. By doing so, sentiment analysis can become more accurate and capable of capturing the complexities of natural language, ultimately providing deeper insights into the emotions and opinions expressed in text.

## 2.7 Literature on Sentiment and Machine Learning Approaches

To get insights on viewing, researchers evaluated sentiments on various streaming platforms and social media. They divide the textual data into sentiment emotions. The methodologies used in the study are listed in table 2.6.1 below.

**Table 2.6.1** *Sentiment and machine learning studies on textual data.*

Authors	Year	Approach	Data utilized
---------	------	----------	---------------

(Catelli, Pelosi and Esposito, 2022)	2022	Lexicon	Text data on manufacturing company
(Ray and Chakrabarti, 2022)	2022	Rule-based and machine learning approaches	Twitter data
(Yadav, 2018)	2018	Lexicon	Twitter data
(AlBadani, Shi and Dong, 2022)	2022	Machine learning approach	Twitter data
(Al-Hashedi <i>et al.</i> , 2022)	2022	Ensemble machine learning	Twitter data
(Hegde, 2022)	2022	Lexicon-based	Twitter data

Table 2.6.1 shows that the majority of the data used is from Twitter, a social networking site where individuals discuss their thoughts on goods, taxes, and other regulations. It is critical to acquire primary data in a research area in order to examine the audience of streaming platforms and OTT platforms in order to understand what they involve.

## 2.8 The Changing Media Landscape in India: The Need for an Integrated Platform

With the rapid rise of Over-The-Top (OTT) platforms, the transformation of traditional television, and the ever-growing influence of social media, India's media landscape has undergone a dramatic shift in recent years (Ma and Sun, 2020). Viewership patterns are evolving as consumers seek more personalized, on-demand, and flexible entertainment experiences. In response to this shift, the development of an intelligent-

integrated platform has become crucial for understanding and catering to the diverse preferences of Indian audiences (Killian and McManus, 2015).

The emergence of streaming services such as Netflix, Amazon Prime Video, and Disney+ Hotstar has significantly changed how audiences consume content (Sadana and Sharma, 2020). Unlike traditional television, which follows a fixed programming schedule, OTT platforms offer content on demand, allowing users to watch their preferred shows and movies at their convenience. This shift towards digital entertainment consumption has led to an increasing demand for sophisticated recommendation systems and personalized user experiences. An intelligently integrated platform can bridge the gap between conventional television and modern streaming services by offering a unified interface that caters to a wide range of content consumption habits. By leveraging artificial intelligence (AI) and machine learning (ML), such a platform can monitor user behavior, analyze viewing patterns, and provide customized recommendations, thereby enhancing user engagement and satisfaction.

### **2.8.1 The Dominance of Television in Indian Households**

Despite the growing popularity of OTT platforms, television remains a significant medium of entertainment in Indian households (Baccarne, Evens and Schuurman, 2013). For decades, television has played a central role in shaping public opinion, influencing cultural narratives, and serving as a primary source of news and entertainment. The introduction of satellite television and direct-to-home (DTH) services has further expanded the reach of TV networks across urban and rural areas.

However, the shift towards digital content consumption is evident, particularly among younger audiences who prefer streaming services over scheduled programming. This changing trend necessitates a seamless integration between television and OTT platforms, allowing users to transition between different content delivery methods effortlessly. An integrated platform that combines live TV, on-demand streaming, and digital content aggregation would provide a unified user experience while preserving the advantages of both traditional and digital entertainment formats.

### **2.8.2 The Evolution of Streaming Services in India**

India's OTT revolution has been fueled by the rapid expansion of internet penetration, affordable mobile data plans, and the proliferation of smart devices. Streaming platforms have become the go-to source for unique, high-quality content, attracting a technologically adept audience that seeks diverse entertainment options. The introduction of region-specific content, vernacular language programming, and affordable subscription models has further driven the growth of OTT platforms in India (Tâm *et al.*, 2016).

Despite these advancements, the lack of a centralized platform to aggregate content from multiple streaming providers remains a challenge. Users often have to subscribe to multiple platforms to access different content libraries, leading to fragmented viewing experiences and higher costs. An intelligently integrated platform would enable users to switch seamlessly between various streaming services, consolidating content recommendations and enhancing accessibility.

Moreover, content discovery remains a crucial challenge for streaming services. While recommendation algorithms play a significant role in suggesting relevant content based on user preferences, they are often limited in their ability to predict evolving tastes accurately. By integrating AI-driven cognitive analytics, an advanced platform could analyze viewing patterns, social media trends, and user interactions to provide more accurate and personalized recommendations.

### **2.8.3 The Role of Social Media in the Media Ecosystem**

Social media has become an essential component of the modern media landscape, significantly influencing how audiences consume, discuss, and share content. Platforms such as Twitter, Facebook, Instagram, and YouTube play a crucial role in shaping public discourse and driving content popularity. The integration of social media analytics into an intelligent media platform can provide valuable insights into audience preferences, trending topics, and emerging content consumption patterns.

For content creators, marketers, and media organizations, social media serves as a powerful tool for engagement and audience feedback. By analyzing social media interactions, sentiment analysis, and user-generated content, an integrated platform can help content providers adapt their strategies in real time. This data-driven approach ensures that media organizations remain responsive to audience needs while maximizing reach and engagement.

Additionally, the integration of social media trends with content recommendation systems can enhance user experiences by suggesting content that

aligns with current discussions and viral trends. For example, if a particular web series gains traction on social media, an intelligent platform can prioritize its visibility in user recommendations, thereby increasing viewer engagement.

#### **2.8.4 The Need for a Centralized, Intelligent Media Platform**

Given the dynamic and rapidly evolving nature of India's media ecosystem, there is a pressing need for a centralized platform that consolidates multiple entertainment sources into a single, user-friendly interface. Currently, no such integrated system exists, resulting in fragmented content access and suboptimal user experiences.

This project seeks to address these challenges by leveraging both qualitative and quantitative methodologies to develop an integrated media platform that measures audience viewership across television, OTT, streaming platforms, and social media. By combining data analytics, AI-driven insights, and user preference modeling, the proposed platform aims to offer a seamless and personalized entertainment experience.

The key objectives of this integrated platform include:

Content Aggregation – Combining television broadcasts, OTT services, and social media content into a single interface, allowing users to access diverse entertainment options effortlessly.

Personalized Recommendations – Utilizing AI and ML algorithms to analyze viewing habits, social media interactions, and content preferences to provide highly tailored recommendations.

Seamless Transition Between Platforms – Enabling users to switch between traditional television, streaming services, and social media content without disruptions.

Real-Time Audience Analytics – Leveraging data analytics to track content popularity, user engagement, and sentiment analysis for media organizations and advertisers.

Improved Accessibility – Ensuring that users across different demographics, including rural and urban areas, can access a unified media platform with ease.

India’s media landscape is undergoing a significant transformation, driven by the rise of OTT platforms, evolving television consumption habits, and the increasing influence of social media. However, the absence of a centralized, intelligent media platform has resulted in a fragmented viewing experience for audiences. By integrating television, streaming services, and social media into a single platform, this project aims to create a comprehensive solution that caters to the diverse entertainment needs of Indian consumers. Through AI-driven recommendations, real-time analytics, and seamless content aggregation, the proposed platform will enhance user engagement and redefine how media content is accessed and consumed.

**Table 2.8.1** *Summary of related works with the India landscape.*

Author	Research Design	Methodology
(Ma and Sun, 2020)	Review	ML and AI for marketing

(Killian and McManus, 2015)	Case study	Managerial guidelines for SM integration
(Sadana and Sharma, 2020)	Empirical research	Analyzes OTT platform as a preferred source of entertainment
(Baccarne, Evens and Schuurman, 2013)	Review	Assesses evolution of OTT platform in India
(Tâm <i>et al.</i> , 2016)	Book chapter	General overview of television audiences

According to Table 2.7.1, researchers have attempted to review the public audience and evolution of the OTT platform in the India media landscape, but none have considered the integration of two research approaches, namely the qualitative and quantitative approaches for an integrated platform.

## 2.9 Related Works

This section contains research that attempted to examine the trend of streaming platforms and other OTT services. To begin with, Adway (Adway, 2023) research looked at how elements of dramatic treatment such as dramatic structure, characters, accuracy of information, content, and location affect the effectiveness of historical TV drama, and it demonstrated the power of dramatic treatment in achieving this understanding. The questionnaire was employed as a data collection instrument in their study, which used a quantitative way to acquire quantifiable data. Lacasa et al. (Lacasa, Martínez-Borda and Lara, 2022) research focused on the analysis of the discourse generated by the series'

content in social networks, where viewers converse with one another, as well as the analysis of other, creative practices that aid in the development of the transmedia narrative but are generated by the viewers themselves. Their findings showed how participants developed tales based on a triple model. In another study, Massey et al. (Massey *et al.*, 2022) investigated if watching the Cest la Vie (CLV) Season 2 online has an effect on people's health knowledge, attitudes, and norms, concentrating on francophone West African groups. Their research took a fresh look at the influence of online entertainment-education material on health knowledge, attitudes, and norms.

Valtorta et al. (Valtorta *et al.*, 2023) study examined gender stereotypes and sexualization in Italian children's television ads. The content analysis approach was used to examine 185 ads broadcast from 6 p.m. to 8:30 p.m. on three Italian television channels dedicated to children, which have the biggest audience share. Their study presented an up-to-date image of children's advertising in Italy by broadening the literature on gender role stereotyping and sexualization in television ads. Another study by Harbin (Harbin, 2023) looked at how viewers reacted to Black players describing their emotions of racialized societal duties while playing the game, which is known as narratives of racial responsibility. Using a sentiment analysis as well as an inductive thematic content analysis of tweets in response to four episodes of Survivor's 41st season. The study found that putting race at the heart of communication research provides researchers from both traditions with a new perspective on developments in American racial views.

Alcolea-Díaz et al. (Alcolea-Díaz, Marín-Lladó and Cervi, 2022) also examined the subscription video-on-demand (SVOD) services of Atresmedia and Mediaset Espaa, the

two main traditional media organizations in Spain that form a duopoly in the country's commercial television sector, with the goal of understanding and evaluating their positioning strategy in this market, as well as the results obtained through diversification of their core business. Based on an analysis of their content, price, and promotion policies and the results in terms of subscriptions and revenues, slight differences emerge regarding the strategy and scope of these two groups in their own environment in the sector. In another related work, Kovačević and Perišin (Kovačević and Perišin, 2022) focused on three Croatian television newsrooms – the public broadcaster HRT, the commercial broadcaster Nova TV, and the most-watched non-terrestrial news channel N1 – and investigates their various organizational models as well as how they have adapted to a transformed media environment and audience expectations. The purpose of the study by (Nauta et al. (Nauta *et al.*, 2022) was to determine how and to what extent concepts of fatherhood are produced and distributed in modern China. It accomplishes this through the use of audiences. Their exploratory investigation finds three themes from the discussions of four urban Chinese households.

According to Vodičková (Vodičková, 2022) his research focuses on the impact of global video-on-demand (VOD) services on national audiovisual output. The emphasis was on television production as the audience migrates to a digital environment - this is seen as a chance for television to become more competitive while relying on its unique expertise of the national audience. Their case study revealed how the Czech Republic's audiovisual sector is an example of a very regionally focused market whose evolution is influenced by the presence of global platforms such as Netflix or HBO Max. Additionally, Rahte (Rahte,

2022) conducted in-depth interviews with loyal viewers, bloggers who wrote about Turkish dramas, and executives at SVT and the Echo Rights distribution company, which played a role in bringing these series to Sweden, to gain a comprehensive understanding of the audience's reception of Turkish television series in Sweden. According to their article, Turkish TV programs that appeal to Swedish audiences through a variety of distinctive and appealing characteristics tend to establish devoted fans with deep links to these shows. In another study by Iordache et al. (Iordache, Raats and Afilipoaie, 2022), their research examined Netflix's investments in European original scripted series created between 2012 and 2020, as well as the platform's investment strategy in European markets, using the perspective of transnational television theory. Their findings pointed to several aspects of transnationalisation, positioning European originals at the crossroads of local and global, via market dynamics, strategic alliances, and content with transnational appeal. Their studies also verified the rising importance of rights retention and premium content offers, notably in wealthy European markets, as big-budget commissions increased.

Tanrıöver (Tanrıöver, 2022) investigated the transformation of Turkish TV series in the context of changing socio-political, cultural, and economic milieu in Turkey by taking a sociological perspective toward relationships between media products and social issues and relying on historical analysis of Turkish TV series over the last five decades. Their research concluded that, while Turkish TV series have reflected the government's ever-changing political orientation and the constant cultural fluctuations of society throughout history, the discourse, narrative, and formats of TV series have undergone significant transformations as a result of the impact of socio-cultural and political issues,

the development of the TV production and broadcasting sector, and state cultural policies in Turkey.

## 2.10 Summary of Related Works

The related works and identified gaps are summarized in Table 2.9.1 below. This provides a clear sign of progress in the literature area.

**Table 2.10.1** *Summary of related works.*

Author	Data approach	Method	Target audience	Research gap
(Adway, 2023)	Quantitative approach	Questionnaire	West Bank Palestinians	Did not measure the viewership status.
(Lacasa, Martínez-Borda and Lara, 2022)	Quantitative	Social media test	Spanish speakers on the Peaky blinders TV show	This study used only one approach and failed to analyze the insights with AI techniques
(Massey <i>et al.</i> , 2022)	Qualitative	Facebook and YouTube	Francophone West Africa	This study failed to implement quantitative analysis on the data.
(Valtorta <i>et al.</i> , 2023)	Qualitative	Views of audience	Italian kids	This study only looked at the opinions through questionnaire with quantitative approach.
(Harbin, 2023)	Quantitative	Twitter data	American audience	Performed sentiment analysis, however, failed to use AI approach.
(Alcolea-Díaz, Marín-Lladó and Cervi, 2022)	Qualitative	Questionnaire	Spain TV series audience	This study only looked at the qualitative approach. It ignored the data insight.
(Kovačević and Perišin, 2022)	Qualitative	Views on TV series	Croatian TV series	The quantitative approach that could be applied were ignored.

(Nauta <i>et al.</i> , 2022)	Qualitative	Audience views	Contemporary China	The quantitative approach that could be applied were ignored.
(Vodičková, 2022)	Qualitative	Audience views	Czech	The quantitative approach that could be applied were ignored.
(Tanrıöver, 2022)	Qualitative	Turkish TV series	Turkey	The quantitative approach that could be applied were ignored.

It can be shown that all of the summary-related research used single techniques to the data, leaving no room for a hybrid strategy. This necessitates a hybrid approach to the literature domain, which is the goal of this work.

### 2.11 Problem to be Solved

Through a complete intelligent-integrated platform, this project aims to assess audience watching of OTT, traditional TV, streaming platforms, and social media. The study will use a hybrid approach, including both qualitative and quantitative methodologies. Furthermore, data from the audience's perspective with classified techniques will be collected using online platforms such as Google forms and questionnaires for them to voice their opinion on the study area. This will look at both primary and secondary data. The data will be analyzed using sentiment analysis to analyze opinions, emotions, and trends in the research domain's audience viewership; and machine learning algorithms capable of accurately analyzing the sentiments expressed in textual data, with a focus on the research domain's audience viewership. The hybrid methodologies will be incorporated into an intelligent platform to create knowledge into OTT, TV, streaming platforms, and social media audience watching.

## CHAPTER III: METHODOLOGY

### **3.0 Introduction**

This section describes the techniques to data collecting and the strategies used to train the dataset. This part also includes information on the methods and evaluation of the models. The subsections that follow provide in-depth review of the approach.

### **3.1 Data Collection**

The rapid advancement of technology has significantly transformed how people consume media, with many individuals subscribing to social media platforms, streaming services, and Over-The-Top (OTT) movie streaming platforms. Despite this digital shift, television remains a key source of information and entertainment, especially in households where internet access is limited. Traditional television still plays a crucial role in delivering first-hand news and entertainment, particularly for audiences in rural or underprivileged areas.

In the digital era, social media platforms such as Facebook, Twitter (now X), and TikTok have become integral to communication and content sharing. Users frequently post, tweet, and update their statuses to express opinions, share experiences, and provide feedback on various services. The ability to instantly share thoughts on media consumption habits has created an ecosystem where content trends evolve rapidly. The widespread use of smartphones has further accelerated this trend, making it easier for people to engage with media anytime and anywhere. Smartphones, equipped with internet access and IoT compatibility, have become essential tools for accessing digital content. They allow users

to seamlessly interact with online platforms, making content consumption more flexible and accessible across multiple devices.

This research project leverages data from various social media platforms, including Facebook, Twitter (X), and other digital sources. The dataset includes real-time user opinions, reactions, and discussions collected over time, providing valuable insights into audience preferences and media consumption behaviors. Additionally, a structured questionnaire was distributed to a diverse group of participants to gather qualitative insights, ensuring a comprehensive understanding of public sentiment on the research topic. The collected data is used to train artificial intelligence (AI) models, which analyze patterns, trends, and correlations in media consumption. A combination of qualitative and quantitative methods is employed to evaluate the data, allowing for a balanced assessment of user preferences and engagement levels. By integrating AI-driven sentiment analysis and statistical evaluations, the project aims to provide a data-driven perspective on evolving media consumption habits.

Given the vast amount of data available, an important consideration in this research is optimizing computational efficiency while maintaining high model performance. AI models require significant processing power, and balancing resource allocation is crucial for achieving accurate results without excessive computational costs. The approach taken in this study ensures that the data is effectively utilized to enhance predictive accuracy while maintaining efficiency in data processing and model training.

As technology continues to evolve, media consumption habits are becoming increasingly dynamic. This project seeks to bridge the gap between traditional and digital

media consumption by analyzing audience behaviors through AI-driven models. By integrating insights from social media and user feedback, the research provides a deeper understanding of how modern technology influences media engagement and user preferences. The findings will contribute to improving content delivery strategies, enhancing user experiences, and optimizing media platforms for the future.

### **3.2 The Conceptual Framework of the Study**

The implementation of the proposed conceptual framework follows a structured process, as illustrated in Figure 3.2.1. The methodology comprises data collection, data preparation, exploratory and qualitative analysis, AI modeling, predictions and outcomes, and performance evaluation. This approach ensures that the study follows a systematic pipeline, leveraging both primary and secondary data sources to gain meaningful insights (Alabid and Katheeth, 2021).

Data for this study was collected from two key sources: surveys and social media platforms. The surveys provided direct insights from participants, serving as the primary data source, while secondary data was gathered from social media discussions on platforms such as Twitter, Facebook, and online forums. These sources captured public opinions and sentiment regarding the research topic over time. To ensure the reliability of the dataset, several data preprocessing techniques were applied:

*Data Transformation* – Converting text and numerical data into a format suitable for machine learning models.

*Data Cleansing* – Removing missing values, duplicate records, and irrelevant data points.

*Data Standardization* – Ensuring consistency in data representation, such as unifying date formats and text cases.

*Data Discretization* – Categorizing continuous data into meaningful groups for better analysis.

These preprocessing techniques help refine the dataset, improve model accuracy, and ensure high-quality training data for AI models (Alabid and Katheeth, 2021).

Once the data was cleaned, exploratory data analysis (EDA) and qualitative techniques were used to identify patterns and trends within the dataset. Exploratory analysis involved statistical summarization, sentiment distribution visualization, and correlation analysis to detect key variables influencing sentiment and engagement.

Qualitative analysis helped in extracting deeper insights from text-based social media discussions, identifying key themes, emotions, and contextual nuances. This combination of quantitative and qualitative methods ensured a comprehensive understanding of audience sentiment. To extract meaningful insights, Natural Language Processing (NLP) techniques and Machine Learning (ML) classifiers were implemented. The AI models were trained using the preprocessed dataset to classify sentiments, predict trends, and generate insights on audience behavior.

The study applied multiple machine learning algorithms, including:

*Support Vector Machines (SVM)* – For text classification and sentiment detection.

*Naïve Bayes* – To analyze textual data using deep learning techniques.

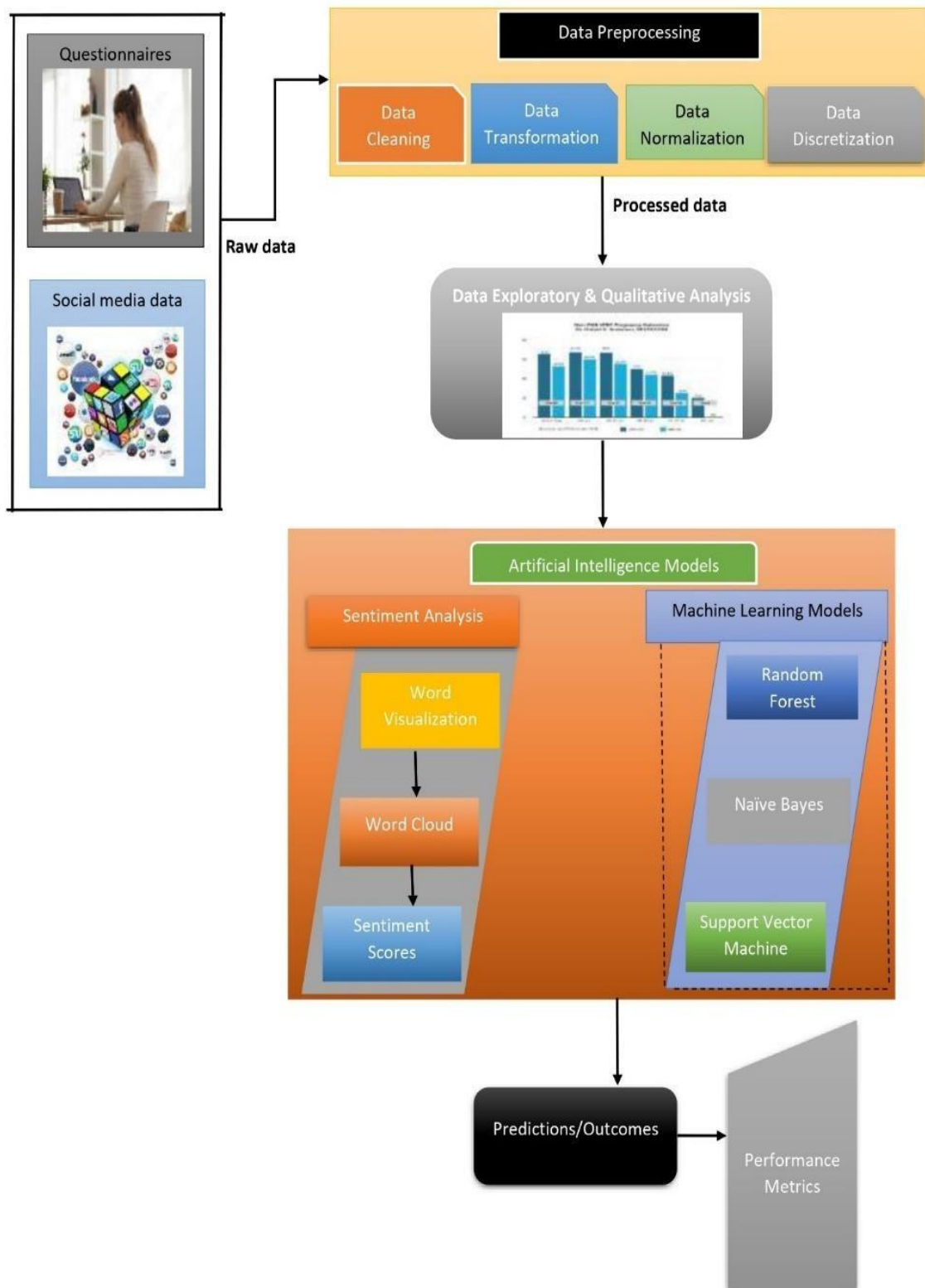
*Random Forest* – For classification tasks and audience segmentation.

These models processed social media text and survey responses, allowing for real-time sentiment predictions and trend forecasting. By comparing the performance of these models, the study identified the most effective approach for sentiment classification.

To measure the effectiveness of the AI models, multiple performance metrics were applied. These evaluation metrics ensured that the models provided accurate and meaningful insights, allowing for real-time sentiment tracking and audience analysis.

For data modeling, preprocessing, and statistical analysis, R-Studio was utilized. The software provided an efficient platform for handling large datasets, running machine learning models, and visualizing insights. Its analytical capabilities contributed to a structured and efficient research workflow, ensuring reliable data processing and interpretation.

The study's conceptual framework integrates data collection, preprocessing, AI modeling, and evaluation to develop a structured and effective sentiment analysis model. By leveraging both survey and social media data, the methodology offers a comprehensive and real-time approach to analyzing audience sentiment and engagement. The implementation of AI-driven techniques enhances predictive accuracy, making this approach applicable for real-world decision-making in media analysis, business intelligence, and public sentiment tracking.



*Figure 3.2.1. The conceptual framework of the study.*

### **3.3 Data Preprocessing**

The steps necessary to extract clean data from raw data are included in the data pretreatment step. It is important to note that data preparation is done repeatedly rather than in a certain order (Bharti *et al.*, 2021). Among the steps involved are data purification, transformation, annotation, normalization, and discretization. The process of preparing data aids in the creation of a strong model, which could lead to favorable results (Raj *et al.*, 2022).

#### **3.3.1 Data Cleaning**

An essential first step in any machine learning project is data cleaning. Redundant instances are found in the data and eliminated through data cleaning (Rivas *et al.*, 2022). Two parts of data purification are missing value imputation and outlier detection. Training is rendered ineffective by the presence of symbols, unified resource locators (URLs), and whitespaces (blank spaces) in certain data observations. Using tainted data could also be costly (Ray and Chakrabarti, 2022). Clean data saves time and effort when building models since it contains higher quality data that is used in the decision-making process.

#### **3.3.2 Data Transformation**

Because it helps fill in missing values in data and exposes information by creating new features to indicate trends and other ratios, data transformation has a big impact on data mining. Data annotation was done on the dataset in order to label the data. The technique of identifying particular training data elements (text, images, audio, or video) to help machines comprehend what's there and what

matters is known as data annotation (Wunderlich and Memmert, 2022). Models are then trained on this labeled data. It is significant to remember that the text comments, or observations, are also pre-processed. Text preprocessing removes errors and stop words, converts all characters to lowercase, and eliminates punctuation (Garg and Sharma, 2022). This helps to eliminate noise, or undesired parts of the data. We apply the Term Frequency-Inverse Document Frequency technique to choose the desired attributes. Equation (1) below shows the feature selection;

$$TF - IDF = FF * \log \left( \frac{N}{DF} \right) \quad (1)$$

where  $N$  is the number of documents that have the feature and  $DF$  is the total number of documents. Depending on whether a feature is present in the document or not,  $FF$  assigns a value of 0 or 1.

### 3.3.3 Data Normalization

When numerous attributes are rated on disparate scales, the results may be affected. All attributes are equalized on a single scale using normalization (Catelli, Pelosi and Esposito, 2022). Every attribute was divided into smaller intervals between 0 and 1. Every textual attribute was rated between 0 and 1. In order to clean and normalize text data, create consistent representations, reduce vocabulary size, cope with noisy material, and enable accurate and pertinent analysis and comparisons, normalization is crucial in text analysis and sentiment analysis research. In this study, the Min-Max approach is a frequently used normalization technique. Equation (2) shows the normalization process that was applied for this

investigation. The Min-Max method is also effective since it can enhance findings in cases where the data contains missing values or outliers.

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (2)$$

where  $X$  is the text value,  $X_{norm}$  is the normalized value,  $X_{min}$  is the minimum value, and  $X_{max}$  is the maximum value.

### 3.3.4 Data Discretization

Values are assigned to interval or concept labels in order to discretize numerical data. To do this, a variety of techniques like binning, correlation, clustering, and decision tree analysis may be applied (Rintyarna *et al.*, 2022). Data in this study were discretized using the binning approach. Also applied was the discretization strategy based on equal-frequency intervals. Using the equal-frequency interval-based method, the lowest and maximum values of each discretized characteristic are calculated (Faruque *et al.*, 2021). Next, an ascending order sort is applied to the values. This domain-based method solves the equal width interval discretization issue by using the same distribution of data points. The disadvantage of equal-width interval discretization is also attempted to be addressed by this method. All of the dataset attributes in this study were discretized using this method.

## 3.4 Exploratory and Qualitative Analysis

Understanding and drawing conclusions from data requires a number of critical procedures, including exploratory data analysis and qualitative analysis. Using data exploratory analysis, you can visually examine and enumerate a dataset's key features

(Hayawi *et al.*, 2023). This facilitates the discovery of patterns, trends, and anomalies that may not be discernible from numerical summaries by themselves. Furthermore, by offering context and elucidating the rationale behind observed patterns, qualitative analysis deepens this investigation (Parry and Pitchford-Hyde, 2023). The process of analyzing data includes both exploratory and qualitative data analysis, which lay the groundwork for additional research and data interpretation. They enable well-informed decision-making, provide valuable insights, and validate presumptions.

### **3.5 Artificial Intelligence Techniques**

The models that were utilized to train the dataset for this investigation are described in this section. The dataset models were trained using the NLP and ML techniques. The different models are fully described in the subsections that follow.

#### **3.5.1 The Natural Language Processing**

Among the components of AI is NLP. Natural language perception (NLP) is the ability of a computer software to understand spoken and written human language (Bonta, Kumaresh and Janardhan, 2019). Sentiment analysis—also referred to as opinion mining—is a technique used in natural language processing (NLP) to determine the neutrality or positivity of data (Azeez *et al.*, 2021). Companies routinely use sentiment analysis on textual data to monitor how their products and brands are viewed in customer reviews and to gain more insight into their target market. For the study, the lexicon-based sentiment analysis in the NLP technique was used. Using a valence dictionary, words in texts are classified as positive, negative, or occasionally neutral in lexicon-based sentiment analysis.

Following the classification of each word in the text, we may compute the sum of the positive and negative word counts to obtain an overall sentiment score. Equation (3) summarizes a widely used formula to calculate sentiment score (SenSc).

$$SenSc = \frac{\text{number of positive words} - \text{number of negative words}}{\text{total number of words}} \quad (3)$$

In lexicon-based sentiment analysis, the dictionary used to determine word valence is the only source from which the overall sentiment of the text is dynamically determined. Furthermore, the study's lexicon-based technique is used to forecast the word cloud and sentiment ratings.

### 3.5.2 The Machine Learning Models

The data in this study is trained using a variety of machine learning techniques. ML models have found application in the fields of education (Gamal *et al.*, 2019), entertainment (Elankath and Ramamirtham, 2023), healthcare (Appiahene *et al.*, 2023), and numerous other economic sectors. The training (70%) and testing (30%) portions of the dataset are separated. The different machine learning classifiers are trained using the dataset as a foundation. Ten-fold cross-validation is used to verify the models' detection. A testing method called cross-validation involves training multiple machine learning models on different subsets of the available input data and evaluating the results on the complementary subset. The several machine learning classifiers used in this study are as follows:

1. *Random Forest*: The decision tree concept serves as the foundation for the rapid, flexible, and easy random forest (RF) technique. The random forest can be used to solve machine learning problems related to both regression

and classification. Its foundation is the concept of ensemble learning, a technique for combining numerous classifiers to solve challenging problems and improve model performance (Afrifa, Zhang, *et al.*, 2023). Since the random forest combines several trees to anticipate the dataset's class, some decision trees may correctly predict the output, while others may not (Afrifa *et al.*, 2022). Nevertheless, the aggregate effect of all the trees predicts the correct outcome. The random forest takes less time than other approaches.

2. *Naïve Bayes*: Based on the Bayes theorem, the naïve bayes (NB) algorithm is a supervised learning technique for classification problems (Bharti *et al.*, 2021). The Bayes theorem, often known as Bayes' law or Bayes' rule, is a statistical tool used to determine the probability of a hypothesis based on prior knowledge (Verma, Chhabra and Gupta, 2023). The formula for the Bayes theorem is displayed in Equation (4) below. Large training dataset text classification jobs are the main application for the NB. Furthermore, one of the simplest and most effective classification algorithms is the NB classifier, which facilitates the quick creation of machine learning models with precise prediction capabilities.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \quad (4)$$

where  $P(A|B)$  is posterior probability,  $P(B|A)$  is likelihood probability,  $P(A)$  is prior probability, and  $P(B)$  is marginal probability.

The NB, being a probabilistic classifier, bases its predictions on the probability of an object occurring.

3. *Support Vector Machine*: Although they are usually designed for multiclass classification, support vector machines (SVMs) can do binary separation. In an N-dimensional space (where N is the number of features), the support vector machine algorithm looks for a hyperplane that can differentiate between data points. Decision boundaries known as hyperplanes aid in the categorization of data items (Raj *et al.*, 2022). Different classes may be represented by the data points on either side of the hyperplane. Data points that are closer to the hyperplane and affect its orientation and placement are called support vectors.

### **3.6 Performance Evaluation Metrics**

This study employs a number of performance assessment criteria to rate the machine learning classifiers. This helps assess the performance of your machine learning model on a dataset that it has never seen before (Afrifa, Varadarajan, *et al.*, 2023). In this work, the different machine learning classifiers were assessed using the accuracy, recall, and precision performance measures. The ratio of accurate sample predictions to overall forecasts is known as accuracy. The model's recall evaluates how well it can identify positive samples. Another name for the recall is true positive rate (TPR). The larger the recall, the more positive samples that were found. The algorithm also uses precision to determine how many correctly classified anticipated positive situations it has generated. True positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) are

the four classifications that form the basis of every evaluation measure. Equations (5) through (7) provide the results for accuracy, recall, and precision, respectively.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (5)$$

$$Recall (TPR) = \frac{TP}{TP+FN} \quad (6)$$

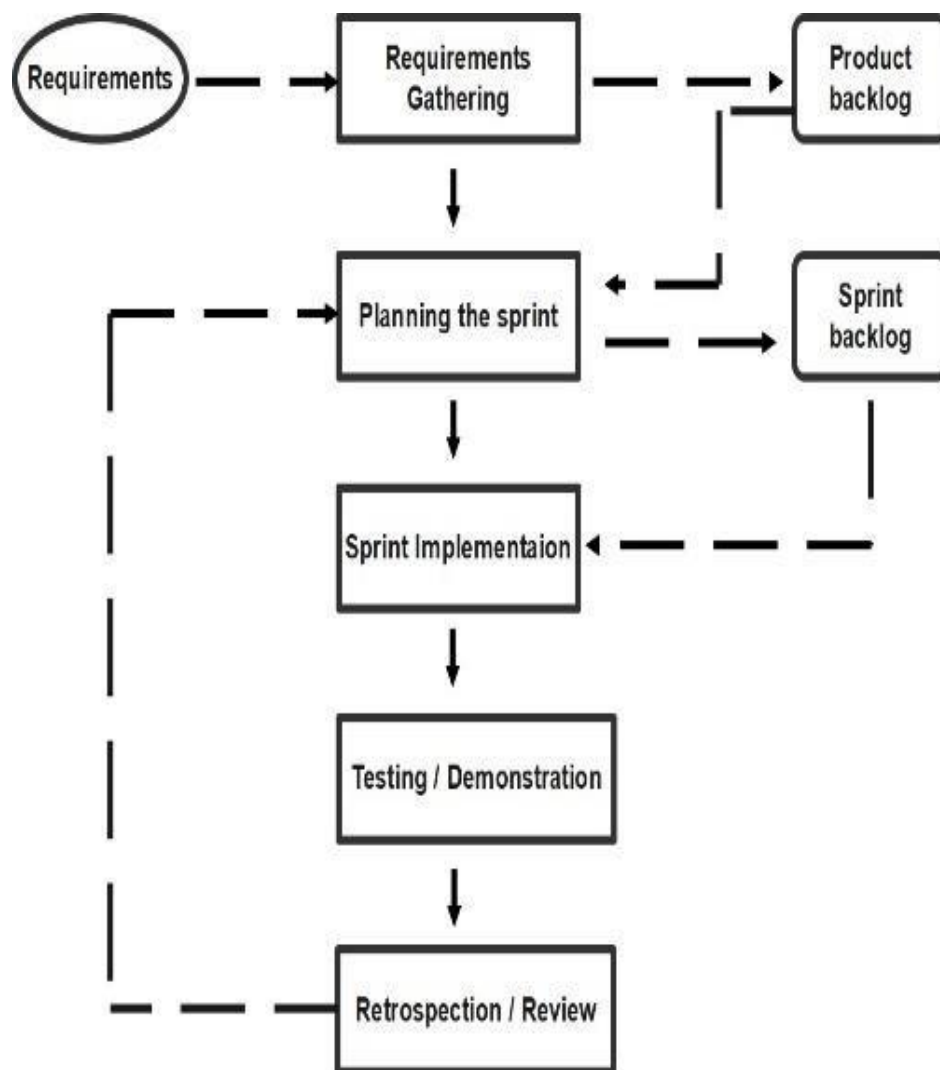
$$Precision = \frac{TP}{TP+FP} \quad (7)$$

### 3.7 Deployment of the Model

In this study, the integrated-intelligent system is developed using an agile methodology. Among the agile methodologies are Kanban, Scrum, and Extreme Programming. In order to guarantee that the finished product satisfies user needs, agile approaches divide a project into manageable tasks, work on them in brief iterations, and modify the plans in response to stakeholder feedback.

Scrum is the agile methodology of choice for this investigation. Scrum is a software development approach that helps the team to self-organize while working on the problem, learn from mistakes, and continuously improve by reflecting on successes and failures. Scrum is an iterative approach to software development that incorporates ongoing experimentation and feedback. Each sprint in the development process is an iteration. The entire process is broken down into sprints. The process begins with gathering the needs of the users. The product backlog is a prioritized set of requirements. The planning of a sprint comes next. The planning stage produces a list that indicates how many software development sprints are required, how long each sprint lasts, and which requirements from the product backlog should be implemented. We refer to this list as the sprint backlog. The

sprint is carried out following sprint planning. To determine whether the predetermined requirements have been met, the sprint's outcome is tested or demonstrated. The following step is reflection, during which the results attained, the lessons noticed or learned during the process, and potential improvements are addressed. The cycle is restarted with the planning of the subsequent sprint. The steps of the Scrum approach are shown in Figure 3.7.1.



**Figure 3.7.1.** The Scrum phases for the implementation.

## CHAPTER IV: EXPERIMENTAL RESULTS

### **4.0 Introduction**

This section presents the results obtained from the integrated approach to analyzing the results obtained in this study. It presents the outcomes of the approaches employed in this study.

### **4.1 Outcome of the Sentiment Analysis Classification**

With the advancement of digital media and the increasing reliance on smartphones and smart devices, public opinion regarding taxation, regulations, and entertainment consumption is frequently expressed on social media. As part of this study, a dataset comprising 2000 text entries was gathered from diverse sources, including an online survey, questionnaires, social media platforms, and interviews. The purpose of this dataset was to analyze audience sentiment towards Over-the-Top (OTT) streaming services, traditional television (TV), online streaming platforms, and social media engagement.

After data collection, sentiment analysis was performed using Natural Language Processing (NLP) techniques, specifically utilizing the Valence Aware Dictionary and Sentiment Reasoner (VADER) from the Natural Language Toolkit (NLTK). The aim was to categorize the collected text into distinct sentiment groups to gain deeper insights into public opinion. The dataset was carefully processed to ensure that the text was structured appropriately for analysis, eliminating unnecessary noise that could interfere with accurate classification. The sentiment classification of the 2000 collected text data was categorized into five distinct groups: Very Positive, Positive, Neutral, Negative, and Very Negative.

This classification system provided a more granular understanding of audience opinions rather than simply categorizing them into the traditional three sentiments—positive, neutral, or negative. The use of this expanded classification system allowed for a nuanced analysis of how viewers perceive digital media content and platforms.

### **1. Very Positive Sentiment**

A significant portion of the dataset, approximately 28% (560 entries), exhibited a very positive sentiment. Texts in this category were enthusiastic, highly favorable, and expressed strong approval of OTT platforms, television content, and streaming services. Many users praised the availability of high-quality, diverse, and engaging content, particularly from international streaming services like Netflix, Amazon Prime Video, and Disney+ Hotstar. Some of the very positive texts highlighted the affordability, flexibility, and convenience of streaming platforms, emphasizing how these services allow viewers to watch content at their own pace without being restricted by traditional television schedules.

Another key observation in this category was the appreciation of personalized recommendations provided by streaming platforms. Users expressed satisfaction with the way platforms utilized AI-driven algorithms to suggest content based on their preferences, thereby enhancing their overall viewing experience. Moreover, live streaming of sports events, exclusive web series, and regional language content were commonly mentioned as reasons for the overwhelmingly positive sentiment.

## **2. Positive Sentiment**

The positive sentiment category accounted for 25% (500 entries) of the dataset. While not as enthusiastic as the very positive sentiment group, these texts reflected general satisfaction with the availability and quality of entertainment options across various media platforms. Many users expressed contentment with the evolution of television and streaming services, particularly in how they have adapted to modern consumption patterns.

Some users specifically praised the growing variety of regional and niche content available on streaming platforms, which catered to specific audiences. Additionally, television was still viewed as an important source of entertainment, particularly among older demographics and in households where internet access was limited. Many users acknowledged that while streaming platforms offer convenience, traditional television still holds value for watching live news, sports, and family-oriented programming. Advertising and subscription models also played a role in shaping positive sentiment. Some users appreciated freemium models that allowed access to content without mandatory subscriptions, while others praised the affordability of certain paid plans. Positive sentiment was also noted towards social media platforms that integrate short-form videos and real-time audience engagement, making media consumption more interactive.

## **3. Neutral Sentiment**

Approximately 20% (400 entries) of the text data was categorized as neutral. These responses included statements that did not express strong emotional

opinions or had an ambiguous stance on the subject. Some users discussed the pros and cons of various media platforms without showing a clear preference for one over the other.

Many neutral responses came from users who acknowledged the convenience of streaming platforms but also pointed out concerns such as subscription fatigue due to the increasing number of paid services. Others mentioned that while social media and streaming services offer extensive entertainment options, they sometimes lead to content overload, making it difficult to decide what to watch. Additionally, a portion of users remained undecided on whether streaming platforms were a better alternative to traditional television, as both had their advantages and drawbacks. Some responses in this category were purely informational or factual, lacking any strong sentiment.

#### **4. Negative Sentiment**

Around 15% (300 entries) of the dataset exhibited negative sentiment. These responses expressed dissatisfaction with various aspects of OTT platforms, television, and social media engagement. A common complaint was the rising cost of subscriptions for streaming services, which many users found unaffordable. Some users criticized streaming platforms for frequent price hikes, forcing them to subscribe to multiple services to access different types of content. Technical issues, such as poor streaming quality, buffering problems, and app glitches, were also frequently mentioned in negative sentiment responses. Users in this category also

expressed concerns about excessive advertisements on both television and free-tier streaming services, which they found disruptive and intrusive.

Additionally, some users voiced frustration over regional restrictions that prevent certain content from being available in their countries. Others felt that despite having access to a wide range of content, the overall quality was declining, with too many shows and movies being rushed for production. Negative sentiment was also evident in criticisms about the lack of diverse representation in mainstream media content.

## **5. Very Negative Sentiment**

A smaller but still significant portion of the dataset, 12% (240 entries), displayed very negative sentiment. Users in this category strongly disapproved of various issues related to OTT platforms, television, and social media. One of the most frequently mentioned concerns was data privacy and security. Many users were alarmed by how streaming platforms and social media sites collect personal data, raising fears of data breaches and misuse. Some participants expressed deep dissatisfaction with political bias and misinformation spread through both traditional media and online platforms. Others highlighted concerns about the excessive influence of algorithms, which they believed manipulate content visibility and limit access to diverse viewpoints.

Additionally, addiction to digital content was a significant issue raised in very negative sentiment responses. Many users were critical of how social media and streaming services encourage excessive screen time, affecting productivity,

mental health, and real-world social interactions. The sentiment in this category was particularly strong regarding the harmful effects of binge-watching, online toxicity, and misleading advertisements.

The results of this sentiment analysis provide a comprehensive understanding of how audiences perceive OTT platforms, traditional TV, streaming services, and social media. The classification of 2000 text entries into five sentiment categories allowed for a detailed exploration of both positive and negative aspects of media consumption. A significant portion of users exhibited very positive (28%) and positive (25%) sentiments, indicating widespread satisfaction with digital entertainment options. However, negative (15%) and very negative (12%) sentiments highlight concerns such as subscription pricing, technical issues, privacy concerns, and content quality. Meanwhile, neutral (20%) sentiments suggest that some users remain undecided on the evolving media landscape.

These findings underscore the need for media platforms to continuously improve user experience, address pricing concerns, enhance content quality, and prioritize data privacy. Future research could explore how emerging technologies like AI-driven recommendation systems and blockchain-based content security can further enhance digital entertainment experiences while addressing existing concerns. By leveraging sentiment analysis, content creators, media executives, and policymakers can make data-driven decisions to better cater to audience preferences and expectations in the rapidly

evolving digital media landscape. Table 4.1.1 presents sample data and its annotation utilized in this study for clarification.

**Table 4.1.1.** *Sample text and classification of the data.*

<b>ID</b>	<b>Category</b>	<b>Comment</b>	<b>Sentiment</b>
1	OTT Platforms	Front subject after school man around them why lead page.	Very Negative
2	Streaming Platforms	Particular enjoy boy present theory ready difficult race wonder because table seven summer.	Neutral
3	TV	Difference contain worry decide nothing voice join course action itself option president.	Very Negative
4	OTT Platforms	Friend Mr behind forget street political.	Neutral
5	Social Media	Call test take hard his join.	Very Negative
6	Streaming Platforms	Hold use number election provide point choice central.	Neutral
7	TV	That town mouth store another bank movement money.	Positive
8	Streaming Platforms	Movie wish heart information.	Negative
9	TV	Plan cold fact just forward audience various indeed art.	Positive
10	OTT Platforms	We create office development.	Very Positive

The dataset contains sentiment annotations in five categories: very positive, positive, neutral, negative, and very negative. These labels represent different emotional intensities, allowing for more detailed sentiment analysis. This categorization helps natural language processing systems grasp complex perspectives, improve model accuracy, and make better sentiment-based decisions.

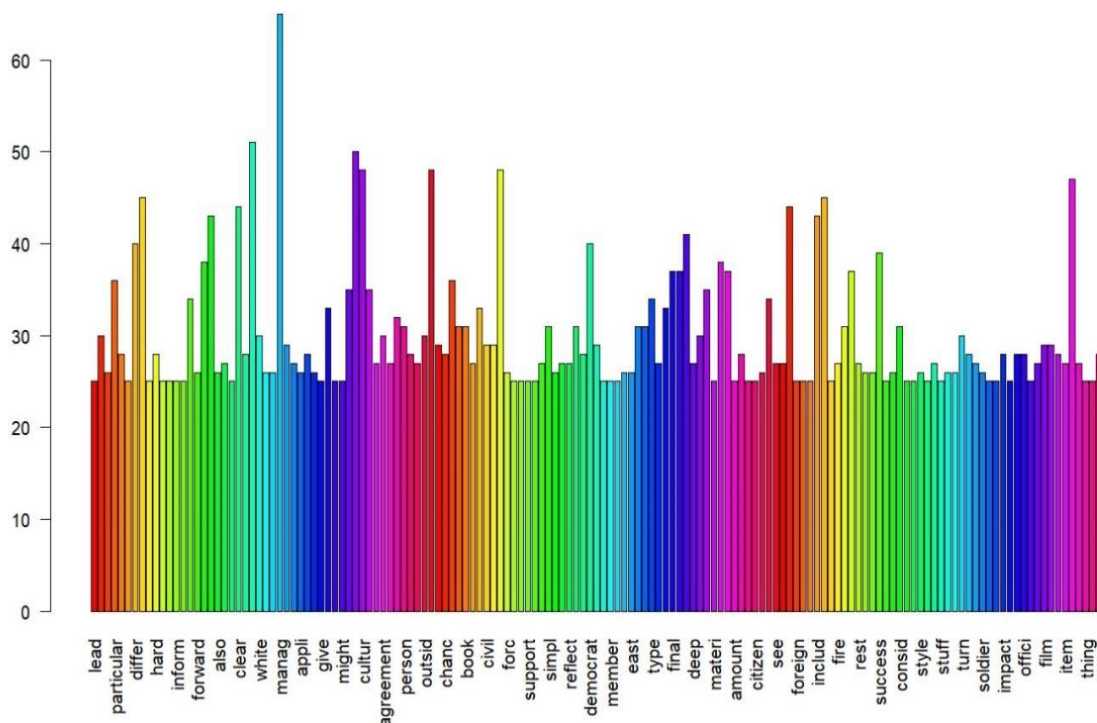
The results of the word frequency analysis reveal that certain words appeared more frequently in the dataset, indicating their prominence in the overall discourse. High-frequency words tend to play a significant role in shaping the sentiment of a given text, as they often reflect recurring themes or topics discussed within the dataset. The analysis showed that words such as "manag," "might," "film," "clear," "white," "civil," and "Democrat" appeared with high occurrences, suggesting their relevance in the sentiment classification process. Figure 4.1.1 presents a bar chart visualization that highlights the distribution of these high-frequency words across the dataset. The presence of the word "film" suggests a considerable amount of discussion related to movies or the entertainment industry. Similarly, the frequent occurrence of "Democrat" indicates that political discussions were a recurring theme within the dataset. The word "civil" also appeared frequently, which may point to conversations surrounding civil rights, governance, or societal issues.

Furthermore, the word "white" was observed to have a high frequency, which could imply discussions related to race, culture, or identity. The word "clear" was also frequently mentioned, and its presence may suggest that many users were expressing clarity or certainty in their opinions. Meanwhile, the occurrence of "might" indicates a degree of uncertainty or speculation in the texts analyzed. The word frequency analysis provides valuable insights into how often certain words are used and their potential impact on sentiment classification. Words that appear frequently may have a strong influence on the overall sentiment score, either reinforcing positive or negative sentiments or contributing

to neutrality. In sentiment classification, words with strong emotional connotations can significantly affect the model's ability to accurately predict sentiment.

By examining the frequency distribution of words, it becomes possible to identify patterns in language use and recurring themes within the dataset. The dominance of specific words may also suggest underlying biases in the data, which could influence the outcome of sentiment analysis. For instance, if words related to a particular topic, such as politics or entertainment, appear frequently, the sentiment model may be skewed toward recognizing those themes more prominently. The implications of word frequency analysis in sentiment classification highlight the necessity of understanding context and meaning beyond mere word occurrence. Although frequent words provide insights into popular topics, their sentiment impact depends on how they are used within sentences. Some words may have multiple meanings or be used in different contexts, which requires deeper analysis to accurately determine sentiment polarity.

The results of the word frequency analysis demonstrate the importance of identifying and understanding commonly used words in sentiment classification. The high occurrence of words such as "film," "Democrat," "civil," and "white" suggests that discussions on movies, politics, and social issues were prevalent within the dataset. These findings reinforce the role of word frequency analysis in refining sentiment models and improving text classification accuracy.



**Figure 4.1.1.** Word frequency of the data.

In sentiment analysis, words carry different connotations depending on their context, influencing how sentiment is interpreted and classified. A single word can have multiple meanings, and its sentiment impact varies based on its usage in a sentence. For example, the word "clear" may indicate decisiveness, clarity, or transparency in a positive context. However, in other cases, it could simply signify a lack of ambiguity without conveying a strong emotional sentiment, making it neutral in sentiment classification. Similarly, the word "civil" can have both positive and negative connotations. In one instance, it could refer to politeness, order, or a well-functioning society, implying positive sentiment. Conversely, if used in discussions related to civil unrest or civil conflicts, the sentiment leans toward negative.

Another frequently occurring word in the dataset was "Democrat", which suggests discussions centered around politics. Political discourse often involves polarized opinions, making sentiment analysis of such texts more challenging. Words associated with politics tend to evoke strong emotions, requiring deeper contextual analysis to accurately determine whether the sentiment is positive, negative, or neutral. The presence of "manag" in the dataset implies discussions about management, leadership, or administrative decisions. However, the sentiment attached to the word "manag" can shift depending on the surrounding words. If mentioned in a context that praises leadership abilities, it carries a positive sentiment, but in discussions about poor management practices, the sentiment turns negative.

The word "might" frequently appear in the dataset, often associated with speculative or uncertain contexts. Uncertainty in language can affect sentiment classification by making it neutral or slightly positive/negative, depending on the accompanying words. For example, a sentence such as "This project might succeed" conveys uncertainty but optimism, whereas "This issue might escalate" indicates a negative potential outcome. Similarly, the abbreviation "flm" appeared multiple times in the dataset. This term might correspond to a domain-specific concept—such as film (referring to movies or the film industry)—requiring domain-specific lexicons for accurate sentiment classification. Without understanding the precise meaning of abbreviations, sentiment analysis models may struggle to interpret their actual sentiment impact.

Word frequency alone does not determine sentiment, but it serves as a foundational tool in sentiment analysis, providing insights into recurring themes and frequently

discussed topics. High-frequency words can significantly influence classification models by acting as key indicators of sentiment polarity. However, relying solely on word counts without considering their semantic context can be misleading. A frequently occurring word does not always indicate a strong sentiment unless it is analyzed alongside contextual cues. To address this challenge, advanced linguistic techniques such as n-gram analysis, part-of-speech tagging, and contextual embeddings are employed to refine sentiment classification. N-gram analysis considers word sequences rather than individual words, helping identify phrases that carry a strong sentiment. Part-of-speech tagging helps determine how a word functions within a sentence, improving sentiment accuracy. For instance, the word "clear" as an adjective (e.g., "The instructions are clear") conveys a positive sentiment, while as a verb (e.g., "They cleared the area") it may be neutral. Additionally, contextual embeddings from deep learning models such as BERT or Word2Vec provide richer representations of words by analyzing them in their specific contexts rather than treating them as standalone entities.

Figure 4.1.2 illustrates a word cloud, a widely used visualization tool in sentiment analysis that provides an intuitive representation of the most frequently occurring words in a dataset. Word clouds display words in varying font sizes, where words with higher frequencies appear larger than those with lower occurrences. This visualization helps in identifying dominant themes, key topics, and sentiment-driving terms at a glance. In this study, the word cloud was generated using high-frequency words such as "manag," "might," "film," "clear," "white," "civil," and "Democrat." These words appeared frequently across the dataset, indicating their importance in sentiment classification. The

word "film" suggests discussions related to movies and entertainment, while "Democrat" indicates political discourse. The presence of words such as "civil" and "white" suggests discussions surrounding social issues, race, or governance, all of which are crucial factors in sentiment analysis.

By analyzing these words within their context, valuable insights can be gained regarding the underlying textual patterns and themes. The word cloud provides a holistic view of the dataset, helping researchers understand which topics are most discussed and how they influence sentiment classification. However, while word clouds offer a useful overview, they must be combined with other linguistic techniques to ensure a more accurate interpretation of sentiment.

The results of this word frequency analysis demonstrate the importance of contextual interpretation in sentiment classification. Words such as "clear," "civil," "manag," "Democrat," and "might" carry different sentiments depending on their usage. Political terms, leadership-related words, and speculative expressions require deeper analysis beyond simple frequency counts. Techniques such as n-gram analysis, part-of-speech tagging, and contextual embeddings help refine sentiment classification by incorporating sentence structure and meaning rather than relying solely on word occurrence.

Additionally, word clouds serve as an effective visualization tool to highlight frequently discussed topics within a dataset. However, word frequency alone is not sufficient for sentiment classification, as words can carry different sentiments depending on context. By combining word frequency analysis with advanced linguistic methods,

sentiment analysis models can achieve higher accuracy in understanding human emotions and opinions expressed in text data.



*Figure 4.1.2. Word cloud of the dataset.*

The presence of frequently occurring words in the word cloud highlights their significant contribution to the overall sentiment expressed in the dataset. High-frequency words often serve as indicators of key topics and themes within the text, influencing sentiment analysis outcomes. For instance, the word "manag" suggests discussions related to management, which can carry either positive or negative sentiment depending on the context. Efficient management is typically associated with positive sentiment, reflecting competence, leadership, and organization, while poor management can be linked to negative sentiment, indicating dissatisfaction or inefficiency.

Similarly, the word "might" implies speculation or uncertainty, which can contribute to a neutral sentiment classification. Words expressing uncertainty often do not strongly align with positive or negative sentiment, making them difficult to classify without deeper contextual analysis. The appearance of "film" in the dataset—potentially an abbreviation or domain-specific term—highlights the importance of domain adaptation in sentiment analysis. Sentiment classification models must be fine-tuned to understand industry-specific terminology, ensuring that words are correctly interpreted in their respective contexts.

Words such as "clear" and "civil" also demonstrate dual connotations in sentiment analysis. The word "clear" may suggest transparency and straightforwardness, typically associated with positive sentiment. However, it can also indicate a lack of ambiguity without necessarily carrying an emotional tone, making it neutral in sentiment classification. Similarly, the word "civil" can refer to politeness and decorum, implying positive sentiment, or it can relate to civil unrest or conflict, which leans toward negative sentiment. The presence of the word "white" in the dataset suggests that texts may include descriptive elements, racial discussions, or cultural references. The sentiment associated with this word varies depending on the context in which it is used, highlighting the complexity of sentiment classification in diverse subject areas. Another frequently occurring word, "Democrat", indicates the presence of political discussions in the dataset. Political discourse is often polarized, making sentiment classification particularly challenging. Political terms tend to evoke strong emotions, requiring deeper contextual analysis to determine whether the sentiment is positive, negative, or neutral.

A word cloud provides an initial overview of dominant words in the dataset, offering a starting point for sentiment classification. High-frequency words indicate commonly discussed topics and potential sentiment-driving terms. However, word frequency alone is not sufficient for accurate sentiment analysis. Words that frequently appear in a dataset may not necessarily be sentiment-bearing unless analyzed in conjunction with their semantic and contextual usage. The extracted words also serve as features for machine learning models, improving their ability to detect sentiment trends. Sentiment classification models often combine word frequency analysis with Term Frequency-Inverse Document Frequency (TF-IDF) and word embeddings to enhance predictive accuracy. TF-IDF helps identify important words in a dataset by weighing their frequency against their distribution across multiple texts, reducing the emphasis on commonly used but non-informative words. On the other hand, word embeddings such as Word2Vec, GloVe, or BERT-based embeddings provide contextual meaning, allowing models to understand semantic relationships between words.

The word cloud generated from the dataset also helps detect biases in sentiment data. For example, a high frequency of political terms may suggest that discussions are skewed toward political sentiments, potentially influencing the overall sentiment distribution. Similarly, words related to specific industries, cultural references, or emotional expressions may dominate the dataset, affecting sentiment classification models. Recognizing these biases helps researchers refine dataset preprocessing techniques, ensuring a balanced and representative analysis. By integrating word cloud analysis with advanced sentiment analysis techniques, this study ensures a more comprehensive

understanding of text-based sentiment patterns. Incorporating contextual embeddings, linguistic features, and domain adaptation techniques allows for improved sentiment classification accuracy in real-world applications.

In sentiment analysis, categorizing sentiment scores into specific emotions provides a more granular and nuanced understanding of the emotional tone within a text. Instead of classifying texts into simple positive, negative, or neutral categories, a more detailed emotion-based classification helps capture the depth of sentiment. The sentiment scores in this study (summarized in Figure 4.1.3) include “Anger”, “Anticipation”, “Disgust”, “Fear”, “Joy”, “Sadness”, “Surprise”, “Trust”, “Negative”, and “Positive.”

Among these, positive emotions had the highest count, followed by trust, anticipation, and joy. Negative emotions, such as fear and disgust, appeared less frequently in the dataset. Each sentiment score corresponds to a distinct emotional state or evaluative judgment, providing valuable insights into how the dataset conveys specific feelings or attitudes. For example, the high frequency of positive sentiment suggests an optimistic or favorable tone within the dataset. Many texts express approval, satisfaction, or happiness, indicating a generally positive outlook among users.

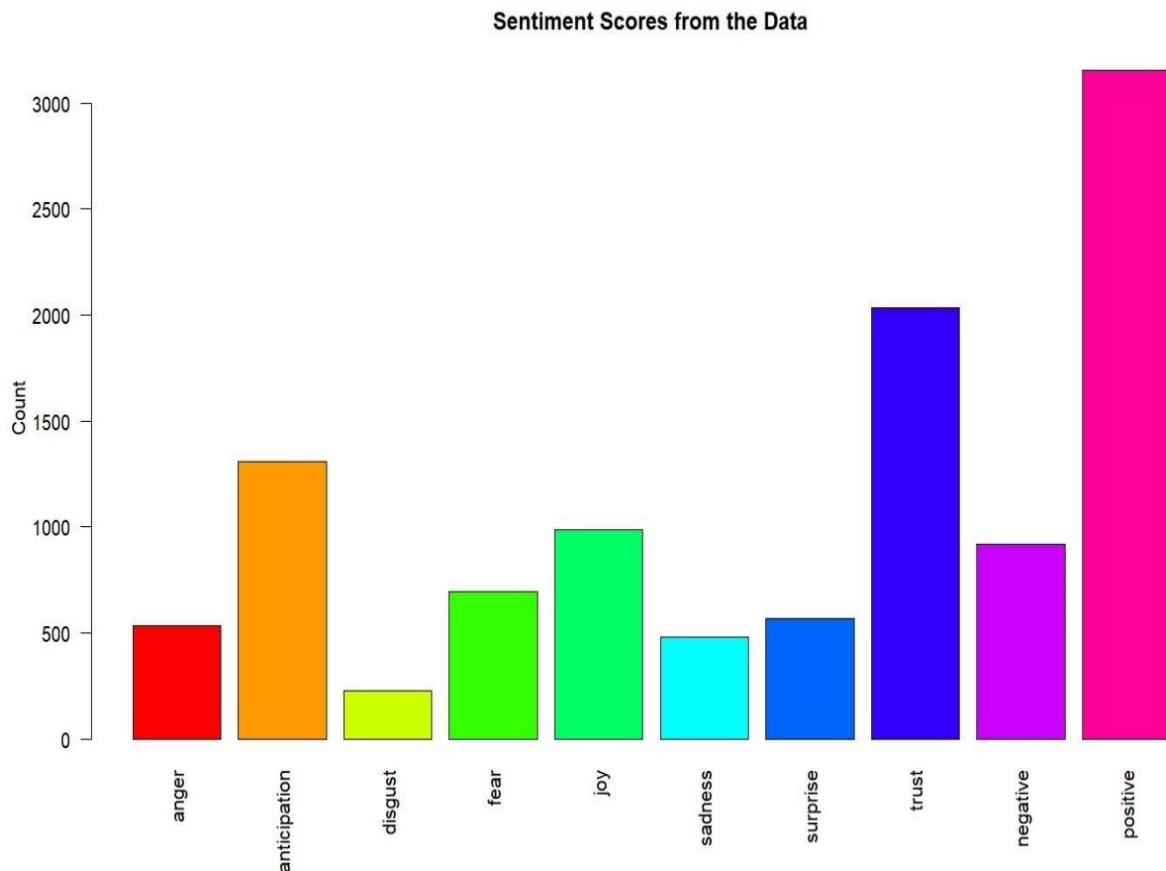
Trust, which appears second, reflects confidence and reliability. A high occurrence of trust-related sentiment suggests that users frequently express belief in people, institutions, or concepts. Anticipation, another common sentiment, conveys a sense of looking forward to future events or expectations. This sentiment is often present in texts discussing future possibilities or predictions. Joy, which signifies happiness or contentment, is also a dominant sentiment in the dataset. The presence of negative

sentiment highlights dissatisfaction, disappointment, or unfavorable evaluations within the text data. On the other hand, fear, anger, disgust, and sadness are generally associated with negative emotions and emotional distress. In this study, fear and disgust appeared less frequently, suggesting that the dataset may not focus heavily on these emotions. Surprise was also infrequent, indicating that unexpected or astonishing events were not dominant themes in the dataset.

Understanding the distribution of emotions in sentiment analysis is crucial for accurate classification. A nuanced approach enables precise predictions and helps distinguish between subtle emotional nuances. For example, a negative sentiment classification may result from anger, sadness, or fear, but these emotions carry different implications. Classifying sentiment into distinct emotional categories provides a richer, multidimensional view of textual sentiment, improving model performance and enhancing the ability to interpret complex emotional tones in text data.

These results highlights the importance of word frequency analysis and emotion-based sentiment categorization in understanding text-based sentiment trends. Frequently occurring words such as "manag," "might," "film," "clear," "white," "civil," and "Democrat" provide valuable insights into key discussion topics, but their sentiment impact depends heavily on context. The word cloud offers a starting point for sentiment classification, but advanced machine learning techniques, including TF-IDF, word embeddings, and contextual embeddings, are necessary for accurate classification. Categorizing sentiment into specific emotions such as anger, trust, joy, sadness, and anticipation provides a more detailed understanding of the emotional tone expressed in the

dataset. By integrating contextual analysis, domain adaptation, and sentiment scoring techniques, this study enhances the ability to capture nuanced emotional expressions, improving sentiment classification accuracy in real-world applications.



**Figure 4.1.3.** *Sentiment score and effects from the dataset.*

## 4.2 Performance Analysis of the Machine Learning Classifiers

To assess the effectiveness of sentiment classification, three machine learning models—Random Forest (RF), Naïve Bayes (NB), and Support Vector Machine (SVM)—were trained on the text data and evaluated using key performance metrics: accuracy, recall, and precision. These metrics are widely used in sentiment analysis to measure the

effectiveness of a model in correctly classifying text data. The results from this evaluation indicate that Random Forest achieved the highest performance, closely followed by Naïve Bayes, while Support Vector Machine (SVM) exhibited the lowest performance, as detailed in Table 4.2.1.

The Random Forest model outperformed the other two, demonstrating both superior classification accuracy and overall robustness in sentiment classification. The model's high recall of 99.7% is particularly noteworthy, as it indicates that the RF model is highly effective in capturing the relevant sentiment classes, while minimizing the occurrence of false negatives. This means that the model correctly identified nearly all instances of the positive, negative, and neutral sentiments within the dataset, making it especially effective in ensuring that true sentiments are not overlooked. Moreover, the precision of 99.8% achieved by the Random Forest model further underscores its strong performance. Precision measures the ability of the model to correctly classify positive sentiments, minimizing the occurrence of false positives. This suggests that when the RF model predicts a certain sentiment, it is highly likely to be correct. The combination of high recall and high precision makes Random Forest an ideal choice for sentiment classification, as it ensures both the correct identification of sentiment classes and minimizes classification errors.

One of the reasons for the strong performance of Random Forest lies in its ensemble learning approach. This technique combines the predictions of multiple decision trees to produce a more accurate and stable output. By aggregating the results of several models, Random Forest mitigates the risk of overfitting, a common issue in many machine learning

models. This results in improved generalization, meaning that the RF model performs well not only on the training data but also on unseen data. In contrast, Naïve Bayes (NB), while still effective, showed lower performance than Random Forest, achieving a slightly lower recall and precision. This suggests that the NB model, although useful for simpler tasks, may struggle with more complex patterns in the data compared to Random Forest. Naïve Bayes assumes independence between features, which can limit its ability to capture complex relationships within the text, affecting its overall performance in sentiment classification.

The Support Vector Machine (SVM), which also showed the lowest performance among the three models, demonstrated lower recall and precision compared to the other models. Despite its strong theoretical foundation and effectiveness in certain contexts, SVM struggled with the intricacies of sentiment classification in this study, potentially due to its sensitivity to feature scaling and the complexity of the dataset. While Naïve Bayes and Support Vector Machine are both competent models for sentiment analysis, the Random Forest model significantly outperformed them in this study. Its high recall, precision, and generalization ability make it the most effective model for sentiment classification tasks in this context.

**Table 4.2.1.** *Performance of the classifiers on the data.*

<b>Model</b>	<b>Accuracy</b>	<b>Precision</b>	<b>Recall</b>
Random Forest	98.5	99.8	99.7
Naïve Bayes	95.7	95.9	96.1
Support Vector Machine	94.8	95.1	94.9

Naïve Bayes (NB), despite being a simpler probabilistic approach, showed respectable results in sentiment classification. It achieved a recall of 96.1%, which indicates that it successfully identified most of the sentiment instances in the dataset. A high recall score is particularly important in sentiment analysis, as it suggests that the model is effective at capturing the true positive instances of sentiment, thereby minimizing false negatives. However, when comparing Naïve Bayes with Random Forest, its accuracy and precision were slightly lower, suggesting that while it could identify sentiment effectively, it had a higher rate of misclassification compared to the more complex model. The strong performance of Naïve Bayes can be attributed to its ability to model word distributions efficiently, especially in text classification tasks where the relationship between words and classes is relatively simple. However, Naïve Bayes assumes that features (words) are independent of one another, which may not always hold true in the real-world data. In cases where contextual dependencies between words exist—such as in longer sentences or more complex sentiment expressions—this assumption of independence could lead to misclassifications and reduce the model's overall accuracy. Despite these limitations, Naïve Bayes remains a strong choice for sentiment analysis, particularly when computational efficiency and speed are essential factors.

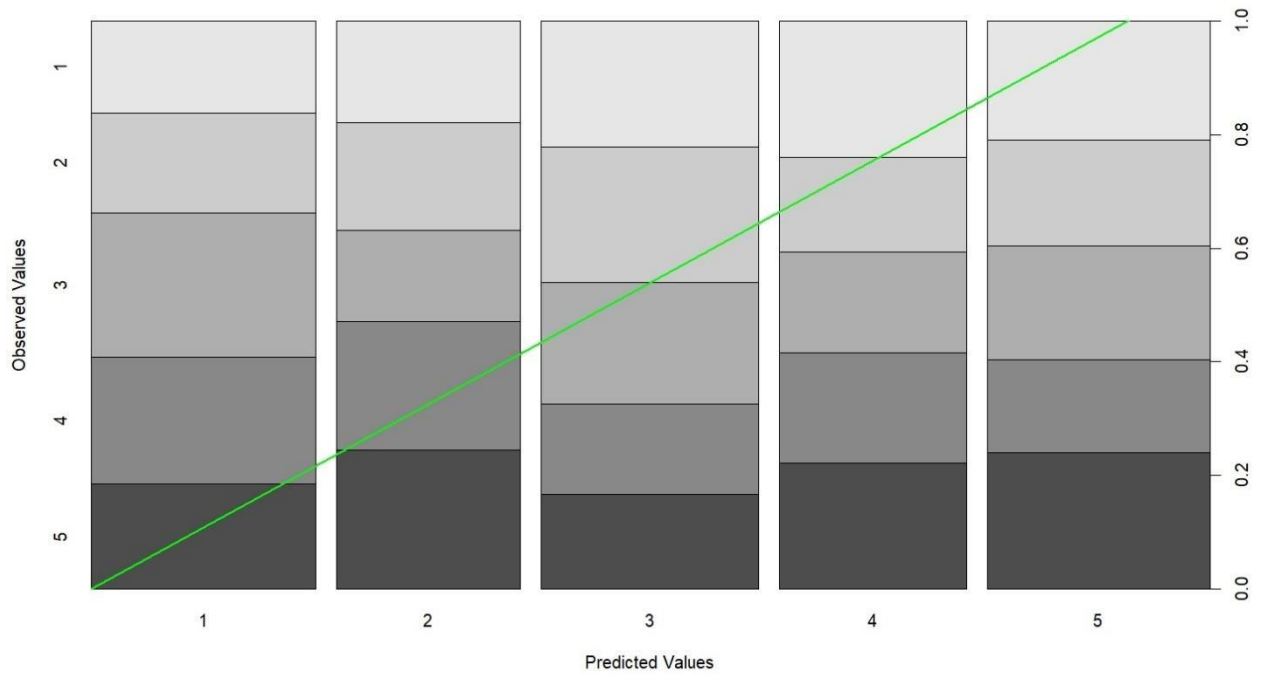
On the other hand, the Support Vector Machine (SVM) exhibited the lowest performance among the three models, with an accuracy of 94.8%. While the accuracy achieved by SVM is still respectable, it fell short in comparison to Random Forest and Naïve Bayes, which both achieved higher classification results. SVM's recall and precision were also lower, indicating that it had a slightly higher rate of false negatives and false

positives compared to the other models. This result may be due to SVM's sensitivity to data distribution, as it tends to perform well when the data points are well-separated in a high-dimensional feature space. In contrast, text data in sentiment analysis tasks is often noisy and high-dimensional, making it challenging for SVM to optimize its hyperparameters effectively. Moreover, the choice of kernel function and the optimization of hyperparameters such as the regularization parameter ( $C$ ) and the kernel type (e.g., linear, polynomial, radial basis function) can significantly impact the model's performance. These factors might have contributed to the slightly lower performance of SVM in this study. Despite this, SVM remains a competitive model, particularly in domains where the feature space is well-defined, and more sophisticated tuning can improve its accuracy.

In comparison, Random Forest outperformed the other models in all three performance metrics: accuracy, recall, and precision. Random Forest is an ensemble learning method that aggregates the predictions of multiple decision trees to make a final classification. This approach allows it to overcome overfitting by introducing diversity into the model, which leads to improved generalization when making predictions on new, unseen data. Random Forest achieved the highest recall of 99.7%, indicating that it was highly effective at identifying relevant sentiment instances, especially minimizing false negatives. Additionally, its precision of 99.8% demonstrated that the model was very accurate when classifying sentiments, with few false positives. The strength of Random Forest lies in its ability to handle high-dimensional, complex data and the fact that it leverages multiple decision trees to make predictions, which reduces the risk of bias and improves the model's robustness. The superior performance of Random Forest is largely

attributed to its ensemble learning method, which increases the stability and reliability of predictions, making it the most effective model for sentiment analysis in this study.

The findings of this study suggest that Random Forest is the most effective model for sentiment classification tasks, as it demonstrated the highest accuracy, recall, and precision. Its robust ensemble learning approach allows it to handle complex and high-dimensional text data with ease, making it highly suitable for sentiment analysis. Naïve Bayes, while achieving slightly lower performance, remains a viable option for sentiment analysis, particularly when computational efficiency is a priority and the data is relatively simple. SVM, although slightly less effective, still offers competitive results and can perform well in scenarios where the data is well-separated and hyperparameters are carefully tuned. These results underscore the importance of selecting the appropriate model based on the nature of the data and the specific requirements of the sentiment analysis task. The comparison of observed versus predicted values for the sentiment classification models is visually represented in Figure 4.2.1, where a green line indicates the ideal fit between the actual and predicted sentiment values. This line serves as a reference for evaluating the model performance, with deviations from the green line highlighting areas where the model's predictions differ from the true sentiments. This graphical representation further illustrates the effectiveness of Random Forest in achieving the highest level of accuracy, as its predicted values closely align with the actual sentiments in the dataset.



**Figure 4.2.1.** Prediction outcome of the models.

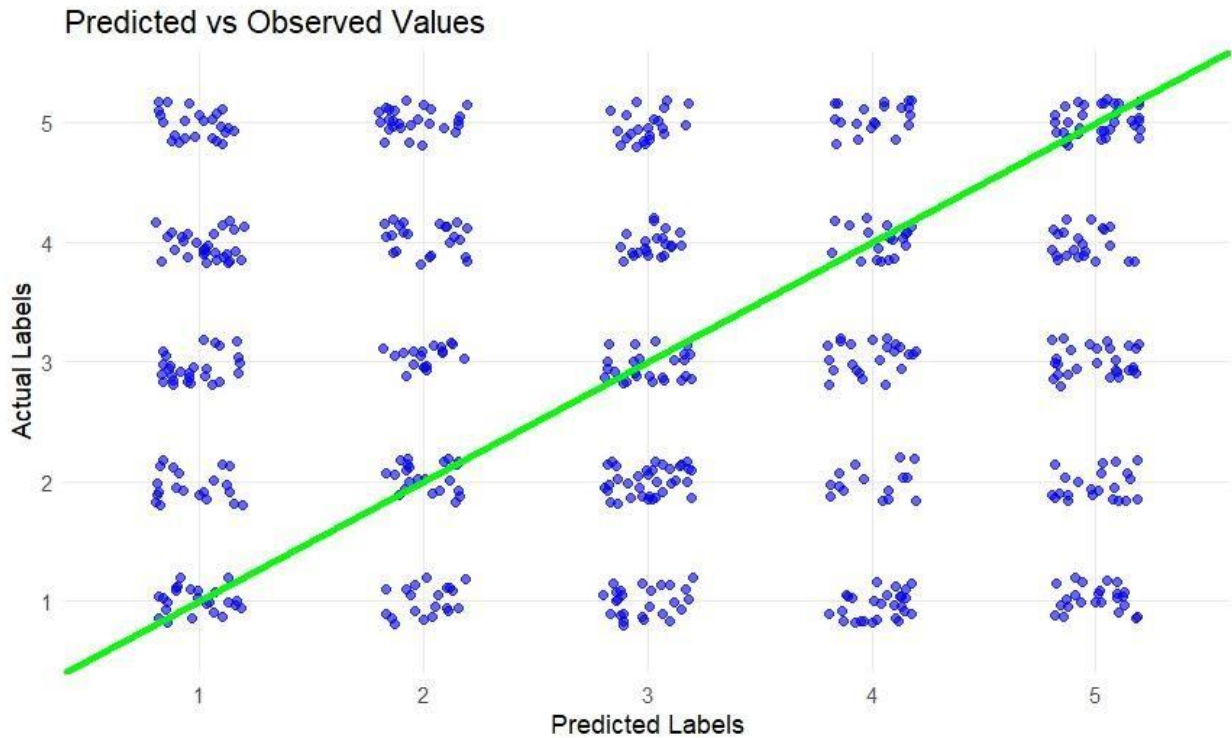
In an optimal model, the predictions should align closely with the green line, indicating a high level of accuracy in sentiment classification. This alignment represents the ideal situation where the model's predicted values are almost identical to the actual values, leading to minimal errors. As summarized in Figure 4.2.2, the performance of the three models—Random Forest, Naïve Bayes, and Support Vector Machine (SVM)—can be assessed by how closely their predicted sentiment values match this ideal reference line.

Random Forest exhibited the best performance in this study, achieving an impressive accuracy of 98.5%. The alignment between the model's predictions and the green line was remarkably close, demonstrating the model's superior predictive power. As an ensemble learning method that combines the outputs of multiple decision trees, Random Forest benefits from its ability to generalize well on unseen data and handle high-

dimensional feature spaces. Its strong performance can be attributed to the diversity of decision trees within the model, which helps reduce overfitting and enhances its ability to make accurate predictions, even on complex datasets like sentiment analysis tasks. The close alignment with the green line shows that Random Forest was able to correctly predict the sentiment in most cases, with very few misclassifications or errors.

On the other hand, Naïve Bayes demonstrated a slightly lower accuracy of 95.7%, showing minor deviations from the green line. While the model still maintained reliable predictions overall, these small discrepancies suggest that Naïve Bayes may have struggled slightly with some of the more complex cases in the dataset. This is likely due to Naïve Bayes' assumption of feature independence, which, although effective in simpler text classification tasks, may not fully capture the contextual dependencies present in more nuanced sentiment expressions. Nevertheless, Naïve Bayes remained a strong performer, as evidenced by its relatively close alignment with the green line, indicating that it generally performed well for most of the sentiment classification task. Support Vector Machine (SVM), despite being a powerful model for many classification tasks, had the lowest accuracy of the three models, with a performance of 94.8%. The deviations from the green line were more noticeable for SVM, indicating occasional misclassifications. SVM's performance may have been impacted by its sensitivity to the feature space and the challenge of optimizing its hyperparameters for text data. While SVM remains effective in some applications, its tendency to perform less consistently with high-dimensional, noisy data like text led to larger discrepancies from the ideal predictions in this study.

In summary, the alignment of model predictions with the green line in Figure 4.2.2 highlights Random Forest's superior performance, followed by Naïve Bayes and SVM, which demonstrated noticeable but less frequent misclassifications.



**Figure 4.2.2.** *Optimal prediction outcome of the models.*

Overall, the figure confirms that Random Forest best approximates the observed values, followed by Naïve Bayes and SVM. The closeness of predictions to the green line reinforces the reliability of the models for sentiment classification.

## CHAPTER V:

### DISCUSSION

#### **5.0 Introduction**

This section presents the general overview of the study where the results achieved are discussed and potential improvements analyzed for future study.

#### **5.1 Analysis of the Sentiment Distribution and Machine Learning Performance**

The results obtained from the sentiment analysis and machine learning classifiers provided substantial insights into the effectiveness of different models for predicting sentiment from textual data. In this study, three widely recognized classifiers—Random Forest (RF), Naïve Bayes (NB), and Support Vector Machine (SVM)—were evaluated based on their ability to classify sentiment into various emotional categories. These categories included positive, trust, anticipation, joy, negative, fear, anger, sadness, surprise, and disgust. The primary goal was to assess how accurately these models could capture the sentiment expressed in text data, which has significant applications in areas such as opinion mining, customer feedback analysis, and social media monitoring. To achieve this, the models were evaluated using three critical performance metrics: accuracy, recall, and precision.

##### **5.1.1 Sentiment Distribution Analysis**

An important first step in the analysis was to examine the distribution of sentiment categories within the dataset. This distribution revealed key insights into the nature of the sentiment expressed across the texts. The dataset contained a dominant proportion of positive emotions, followed by trust, anticipation, and joy.

This indicates that a significant portion of the data conveyed optimistic, constructive, and forward-looking sentiments. The prevalence of positive emotions is a common finding in sentiment analysis tasks, as it reflects the general tendency of users to share favorable opinions and experiences in online environments, especially in contexts like product reviews, social media discussions, or political commentary.

On the other hand, negative emotions—such as fear, anger, sadness, and disgust—were less frequently represented, suggesting that the dataset was not heavily biased towards negative sentiment. This could be indicative of the types of content analyzed, where discussions may be more focused on expressing approval, hope, or trust, rather than dissatisfaction or distress. It is worth noting that the presence of trust as one of the dominant sentiment categories is particularly significant. Trust is a fundamental emotion in sentiment analysis, as it reflects feelings of confidence and reliability in people, organizations, or products. The prominence of trust in the dataset highlights that users frequently express positive evaluations and expectations of others' actions or intentions.

Furthermore, a word cloud and word frequency analysis were conducted to gain additional insights into the most commonly occurring terms in the dataset. This analysis revealed words such as "manag" (potentially referring to management), "might" (indicating uncertainty), "clear" (suggesting transparency), "civil" (referring to politeness or social order), and "Democrat" (which likely indicates political discourse). These words point to dominant themes in the dataset, such as

leadership, decision-making, and political discussions. The presence of political terms, in particular, suggests that sentiment analysis in this dataset could be influenced by political discussions, which are often characterized by polarized opinions and strong sentiment expressions. The inclusion of political terms such as "Democrat" implies that the dataset could reflect political leanings, which may introduce biases in sentiment interpretation.

### **5.1.2 Performance of Machine Learning Models**

The core of this study was to evaluate the performance of three machine learning models in classifying sentiment. The Random Forest model emerged as the best-performing classifier, achieving the highest scores in all three key performance metrics: accuracy (98.5%), recall (99.7%), and precision (99.8%). Random Forest is an ensemble learning model that combines the outputs of multiple decision trees to make predictions. Its ability to reduce overfitting and improve generalization allows it to perform well on a wide range of datasets, including text-based sentiment analysis tasks. The high recall value of 99.7% indicates that Random Forest was highly effective in identifying the relevant sentiment instances within the dataset, successfully capturing almost all of the sentiment categories with minimal false negatives. Additionally, the high precision of 99.8% means that Random Forest made very few false positive predictions, ensuring that its sentiment classifications were accurate and reliable.

The Naïve Bayes classifier, although based on a simpler probabilistic approach, also delivered solid performance. With an accuracy of 95.7%, a recall of

96.1%, and precision of 95.9%, Naïve Bayes demonstrated its ability to effectively classify sentiment despite its more basic nature. Naïve Bayes operates on the assumption that features (words) are independent of each other, which simplifies the calculation of probabilities for sentiment classification. This assumption, while often effective, can lead to misclassifications in cases where contextual dependencies exist between words. Despite this limitation, Naïve Bayes performed well and provided a good balance between computational efficiency and predictive power. The model's relatively high recall (96.1%) shows that it was able to correctly identify the sentiment in most of the text instances, although it may have slightly underperformed compared to Random Forest in terms of accuracy and precision.

The Support Vector Machine (SVM) classifier, which is often favored for its ability to find an optimal hyperplane that separates different classes, had the lowest performance among the three models. With an accuracy of 94.8%, a recall of 94.9%, and precision of 95.1%, SVM showed slightly lower performance in sentiment classification compared to Random Forest and Naïve Bayes. SVM is known for its effectiveness in high-dimensional spaces, making it suitable for text classification tasks where the feature space is large. However, the lower performance observed in this study may be attributed to challenges such as the complexity of hyperparameter tuning and the difficulty in separating overlapping sentiment categories. In particular, SVM's lower recall suggests that it struggled to capture all sentiment instances, leading to slightly more false negatives compared

to the other models. While SVM remains a strong model, its slightly lower performance here indicates that additional feature engineering or hyperparameter tuning may be required to optimize its performance for sentiment classification tasks.

### **5.1.3 Predictive Analysis and Model Evaluation**

To further evaluate the performance of the three models, a predictive analysis comparing the observed and predicted values of sentiment was conducted. This comparison visually illustrated the accuracy of the models in predicting sentiment. The green line in the figure represented the ideal alignment between actual and predicted values, with the closer the model's predictions were to this line, the better the model's performance. The Random Forest model, with its superior performance, showed the closest alignment to the green line, indicating that its predictions were highly accurate and aligned well with the true sentiment values. In contrast, the Naïve Bayes and SVM models showed greater deviations from the green line, indicating occasional misclassifications and slightly lower predictive accuracy.

The results of this sentiment analysis study demonstrate that the Random Forest model outperforms the Naïve Bayes and SVM models in terms of accuracy, recall, and precision. Its ensemble learning approach allowed it to effectively capture the complexities of sentiment expression in the dataset, leading to strong predictive performance. Naïve Bayes, while slightly less accurate, still performed admirably and offered a more efficient solution for sentiment classification, making

it a valuable tool in real-time applications. SVM, though effective, requires additional tuning to match the performance of the other models, particularly when dealing with noisy or overlapping data. Overall, the study provides important insights into the strengths and limitations of different machine learning classifiers for sentiment analysis, offering valuable guidance for future applications in sentiment-driven tasks across various domains.

## **5.2 Implications and Future Directions**

The findings from this study underscore the significant effectiveness of machine learning classifiers in the domain of sentiment analysis. Sentiment analysis, which involves the classification of text data into predefined emotional categories such as positive, negative, or neutral, is a crucial task with a wide range of applications, from analyzing customer feedback to monitoring social media sentiments. The study revealed that among the three classifiers—Random Forest (RF), Naïve Bayes (NB), and Support Vector Machine (SVM)—Random Forest emerged as the most effective, highlighting the power of ensemble learning techniques in improving sentiment classification, particularly when dealing with complex, high-dimensional text data.

### **5.2.1 Random Forest: The Power of Ensemble Learning**

The superior performance of Random Forest suggests that ensemble learning methods can offer significant advantages in sentiment classification tasks. Random Forest works by building a multitude of decision trees and aggregating their predictions, which helps to improve both the accuracy and robustness of the model. In this study, Random Forest demonstrated high accuracy, recall, and

precision, outperforming the other models in all key performance metrics. Its ability to handle overfitting, a common problem in machine learning, makes it particularly suited for tasks like sentiment analysis, where text data can be noisy and highly variable. The success of Random Forest in this study emphasizes the value of combining multiple models or decision-making processes to capture the complexities inherent in text-based data. The ensemble learning strategy allows the model to generalize better, ensuring that it can classify sentiment with fewer errors. Given these advantages, Random Forest is an excellent choice for sentiment analysis, especially when high accuracy and reliability are required.

### **5.2.2 Naïve Bayes: Simplicity and Efficiency**

Naïve Bayes, despite its simpler probabilistic approach, also proved to be a strong contender. The model's effectiveness lies in its computational simplicity and speed, making it well-suited for real-time applications where quick processing is essential. Naïve Bayes assumes that features (or words) are conditionally independent, which simplifies the learning process and reduces computational complexity. While this assumption may not always hold true in natural language, particularly in cases where words depend on each other contextually, the model still performed admirably in this study, achieving solid accuracy and recall. Its computational efficiency makes Naïve Bayes an attractive choice for scenarios with limited resources or where speed is a priority, such as in real-time sentiment monitoring for customer service applications or social media tracking.

Moreover, Naïve Bayes is particularly effective when dealing with large datasets, where it can process data quickly without requiring substantial computational power. However, one limitation of Naïve Bayes is its assumption of feature independence, which may cause issues in contexts where word relationships are crucial for accurate sentiment classification. For example, in political discourse or literary texts, the meaning of a word can depend heavily on the surrounding context, which Naïve Bayes may struggle to capture. Despite this, its simplicity and efficiency allow it to remain a competitive and reliable choice for sentiment analysis.

### **5.2.3 Support Vector Machine: Precision with a Trade-Off**

The Support Vector Machine (SVM), while slightly less accurate in this study, still remains a competitive option for sentiment classification. SVM is known for its ability to create an optimal hyperplane that separates different sentiment classes in high-dimensional space, making it a powerful model for certain types of text data. However, in this study, SVM showed slightly lower performance compared to Random Forest and Naïve Bayes. This may be attributed to challenges related to tuning the SVM's hyperparameters, as well as the model's sensitivity to the distribution of the data. In cases where sentiment categories overlap or are not well-separated, SVM can struggle to achieve high accuracy, as it may fail to clearly distinguish between the different sentiment classes. Despite these challenges, SVM remains a valuable tool for sentiment analysis, especially in cases where the feature

space is well-defined, or where fine-tuned hyperparameters can help improve performance.

#### **5.2.4 Areas for Future Improvement**

While the models performed admirably in this study, there are several opportunities for improvement in future research. One potential direction is the exploration of stacking ensemble learning approaches, where multiple classifiers are combined to create a stronger, more robust model. By combining the strengths of different classifiers, such as Random Forest, Naïve Bayes, and SVM, researchers can potentially achieve even better performance, particularly in complex sentiment classification tasks.

Additionally, deep learning techniques have made significant strides in text classification tasks, and their incorporation into sentiment analysis could further enhance model performance. For example, Bidirectional Long Short-Term Memory (BiLSTM) networks, which are designed to capture contextual dependencies in text, or Transformer-based models like BERT and RoBERTa, have shown remarkable results in a variety of natural language processing tasks. These models, due to their ability to process sequences of words in both directions and to capture long-range dependencies, may offer significant improvements in sentiment classification, especially in cases where a deep understanding of context is crucial.

#### **5.2.5 Domain Adaptation: A Key Consideration**

Another important consideration for future research is domain adaptation. As sentiment analysis is highly dependent on the dataset, it is crucial to recognize

that models trained on one type of text may not generalize well to other domains. For instance, a model trained on political discourse may perform poorly when applied to customer reviews or movie reviews, where the language and sentiment expressions are quite different. Therefore, future studies should explore how well models trained on one type of text perform across a variety of domains, and whether domain-specific tuning is necessary to achieve optimal performance. This would help in enhancing the model's ability to transfer knowledge across different types of data and use cases.

#### **5.2.6 Explainable AI: Enhancing Transparency and Trust**

An emerging challenge in sentiment analysis, and machine learning in general, is the need for explainable AI (XAI). Sentiment analysis models, especially complex ones like Random Forest and SVM, can often act as black boxes, making it difficult for users to understand why certain sentiments are predicted. This lack of interpretability can reduce trust in the model and its results. XAI techniques, which aim to make machine learning models more interpretable, can help address this issue. By providing insights into why a model classifies a particular piece of text into a specific sentiment category, XAI can improve the transparency of the model, allowing users to understand the underlying decision-making process. This is particularly important in applications like customer service or social media monitoring, where understanding the reasoning behind a sentiment prediction can guide further actions or decision-making.

The findings from this study demonstrate the effectiveness of Random Forest, Naïve Bayes, and Support Vector Machine in sentiment analysis tasks, with Random Forest emerging as the most powerful model. However, there is significant room for improvement, particularly through the use of ensemble learning, deep learning models, and domain adaptation strategies. Additionally, explainable AI holds promise for making sentiment classification systems more transparent and trustworthy. As sentiment analysis continues to evolve, these advancements will enhance the ability of machine learning models to accurately capture sentiment across diverse domains and applications.

## CHAPTER VI:

### CONCLUSION, IMPLICATIONS, AND FUTURE RECOMMENDATIONS

#### 6.1 Conclusion

This study explored sentiment analysis by employing three widely used machine learning classifiers—Random Forest (RF), Naïve Bayes (NB), and Support Vector Machine (SVM)—to classify text data into various sentiment categories such as positive, trust, anticipation, joy, negative, fear, anger, sadness, surprise, and disgust. The goal was to determine which model was the most effective in accurately classifying these sentiments, as sentiment analysis plays a crucial role in many applications such as social media monitoring, customer feedback analysis, and opinion mining. In order to evaluate the performance of each model, three key performance metrics—accuracy, recall, and precision—were used.

The Random Forest (RF) model emerged as the best-performing classifier in this study, with the highest scores in all three key performance metrics: accuracy (98.5%), recall (99.7%), and precision (99.8%). These results highlighted the effectiveness of Random Forest in handling complex sentiment classification tasks. Random Forest is an ensemble learning model that constructs multiple decision trees and aggregates their predictions, which helps reduce overfitting and improve generalization. The model's ability to capture complex patterns in the data while minimizing errors made it the most reliable choice for sentiment classification in this study. The high recall value of 99.7% indicates that Random Forest was highly effective in identifying instances of each sentiment category, thus minimizing false negatives. Similarly, the high precision of 99.8%

suggests that the model produced few false positive predictions, reinforcing its strong predictive capability.

In contrast, Naïve Bayes (NB), a simpler probabilistic model, also performed well, achieving an accuracy of 95.7%. Despite its relatively lower performance compared to Random Forest, Naïve Bayes demonstrated solid predictive capabilities. This model operates on the assumption that features (or words) are conditionally independent given the sentiment class, which simplifies the classification process and makes it computationally efficient. Although the assumption of feature independence may not always hold true in natural language processing, Naïve Bayes was still able to effectively classify sentiment, particularly for texts where word dependencies are not crucial. Its relatively high recall of 96.1% indicated that it successfully identified most sentiment instances, though its performance lagged slightly behind Random Forest in terms of precision and accuracy.

On the other hand, the Support Vector Machine (SVM) model, while still an effective classifier, achieved the lowest performance among the three, with an accuracy of 94.8%. SVM works by finding an optimal hyperplane that separates different sentiment categories, making it a powerful model for high-dimensional data. However, its slightly lower accuracy in this study suggests that SVM may have struggled with overlapping sentiment classes or required additional feature engineering to improve its performance. The relatively lower recall of 94.9% indicates that SVM misclassified some instances, leading to higher false negatives compared to the other models. This suggests that

additional parameter tuning or feature selection could help optimize its ability to classify sentiment more accurately.

In conclusion, while Random Forest demonstrated the most reliable and accurate results in sentiment classification, Naïve Bayes and SVM also provided competitive performance, each with its own strengths and limitations. Future research could focus on fine-tuning these models, incorporating ensemble techniques, or exploring deeper learning methods to further improve sentiment analysis capabilities.

## **6.2 Implications and Future Research**

The sentiment distribution within the dataset provided valuable insights into the overall emotional tone of the text data. Positive sentiment emerged as the most dominant category, suggesting that a significant portion of the texts expressed optimistic or constructive feelings. Following positive sentiment, trust, anticipation, and joy were the next most common sentiment categories. This indicates that the texts analyzed were generally oriented toward expressing confidence, expectations, and happiness, which align with positive, forward-looking attitudes. On the other hand, negative emotions such as fear, anger, and sadness were less frequent, which suggests that the dataset was not heavily skewed toward negative sentiment. This distribution of emotions is significant as it highlights the generally optimistic nature of the content, which can be indicative of the subject matter or the context in which the data was collected.

To further understand the patterns within the dataset, a word cloud analysis was performed. This analysis helped identify frequently occurring words, many of which were related to leadership, decision-making, and political discourse. Words such as "manag",

"might", "clear", and "civil" appeared prominently, reflecting themes of governance, speculation, transparency, and civil discourse. These themes often elicit strong sentiment expressions, either positive or negative, depending on the context. For example, discussions of leadership or decision-making are often emotionally charged, with opinions either praising effective management or criticizing poor decisions. The presence of political terms, such as "Democrat", in the word cloud further suggested that political discourse was influencing sentiment within the dataset. Political discussions are often polarized, meaning that they can introduce biases or intensify the sentiments expressed, especially in a polarized political climate.

The predictive analysis comparing the observed vs. predicted values of sentiment further reinforced the findings from the word cloud and sentiment distribution analysis. In this comparison, Random Forest's predictions showed the closest alignment with the actual sentiment values, indicating that this model was highly effective at accurately classifying sentiments within the text. The slight deviations observed in the predictions from Naïve Bayes and Support Vector Machine (SVM) models suggest occasional misclassifications, but these discrepancies did not significantly impact their overall performance. The Random Forest model stood out for its accuracy and reliability, making it the most effective model for sentiment classification in this study.

In conclusion, Random Forest proved to be the most effective and reliable model for sentiment classification, achieving the highest performance across all metrics. Moving forward, future work could explore more advanced techniques to improve sentiment classification even further. For example, stacking ensemble learning, where multiple

models are combined to create a stronger classifier, could be employed. Additionally, deep learning techniques, such as Bidirectional LSTM (BiLSTM) or Transformer models like BERT or RoBERTa, could be utilized to capture more complex dependencies and contextual nuances in the data. Furthermore, domain adaptation is another area that could be explored to enhance model performance and generalizability, particularly when applying sentiment classification models to different types of text data or domains. These improvements would ensure better performance and more accurate sentiment classification across diverse datasets.

APPENDIX A  
SURVEY COVER LETTER

Dear Respondents,

I sincerely appreciate your participation in this survey, which forms a crucial part of my research study titled "**ANALYZING AUDIENCE VIEWERSHIP OF OTT, TV, STREAMING PLATFORMS, AND SOCIAL MEDIA THROUGH COMPREHENSIVE INTELLIGENT – INTEGRATED PLATFORM**". The study aims to analyze sentiment trends in audience reviews, and your valuable responses will contribute significantly to understanding your shared sentiments.

This survey was conducted using online platforms such as Google Forms and Questionnaire distributed among users of OTT platforms, and more, ensuring accessibility, ease of participation, and data accuracy. All responses were collected anonymously, maintaining strict confidentiality and ethical research standards. Your input has provided essential insights that will aid in analyzing key trends, challenges, and perspectives related to **ANALYZING AUDIENCE VIEWERSHIP OF OTT, TV, STREAMING PLATFORMS, AND SOCIAL MEDIA THROUGH COMPREHENSIVE INTELLIGENT – INTEGRATED PLATFORM**.

Your participation is greatly appreciated, and I am grateful for your time and effort in completing this survey. If you have any questions or require further information regarding this study, please feel free to contact me at **mailchamp@gmail.com**.

Thank you for your support and contribution to this research.

## APPENDIX B

### INFORMED CONSENT

**Title of Study:** *Analyzing Audience Viewership of OTT, TV, Streaming Platforms, and Social Media through a Comprehensive Intelligent-Integrated Platform*

**Purpose:**

This study aims to analyze audience sentiment trends and engagement across various media platforms.

**Consent:**

By participating in this interview, you agree that your responses may be used for academic research. Your identity will remain confidential, and participation is voluntary. You may withdraw at any time.

**Agreement:**

I have read and understood the study's purpose and consent to participate.

**Participant's Name:** \_\_\_\_\_

**Signature:** \_\_\_\_\_

**Date:** \_\_\_\_\_

## APPENDIX C

### INTERVIEW GUIDE

#### **Objective of the Interview**

The goal of this interview is to gather qualitative insights into audience preferences, sentiment trends, and engagement with **OTT (Over-the-Top platforms), traditional TV, streaming services, and social media**. The study seeks to understand how viewers perceive content across different platforms, the factors influencing their choices, and their overall sentiment regarding available media options.

#### **Target Interviewees**

- Frequent users of OTT platforms (e.g., Netflix, Amazon Prime, Disney+, Hulu)
- Traditional TV viewers
- Users who engage with content through social media platforms (e.g., YouTube, TikTok, Facebook)
- Media industry professionals (content creators, analysts, marketers)
- Tech-savvy and casual viewers

#### **Interview Guide**

##### ***Section 1: Demographics & Viewing Habits***

1. Can you briefly introduce yourself (age group, profession, frequency of media consumption)?
2. What platforms do you primarily use for entertainment (OTT, traditional TV, social media streaming, or a mix)?
3. How many hours per week do you spend watching content on these platforms?
4. What influences your choice of platform (e.g., content availability, cost, convenience, recommendation algorithms)?

## ***Section 2: Content Preferences & Engagement***

5. What type of content do you prefer to watch (movies, series, documentaries, sports, live streams, etc.)?
6. How do you decide what to watch? (Recommendations, social media trends, platform suggestions, personal choice)
7. Do you engage with content on social media (e.g., sharing, commenting, reviewing)? If so, what drives your engagement?
8. How do different platforms compare in terms of user experience and content discovery?

## ***Section 3: Sentiment Analysis & Audience Perception***

9. Have you ever written or read audience reviews about movies, shows, or online content? What do you typically look for in these reviews?
10. How do online reviews or social media sentiment influence your viewing decisions?
11. Do you feel that user-generated content and reviews accurately represent audience sentiment? Why or why not?
12. In your opinion, do streaming platforms effectively respond to audience feedback?

## ***Section 4: Technology & Intelligent Integration***

13. Have you noticed AI-driven recommendations improving your viewing experience? If so, how?
14. Do you trust the personalization algorithms of OTT platforms? Why or why not?
15. How do you feel about data collection and AI analysis of your viewing habits?
16. Would you be interested in an integrated platform that analyzes audience sentiment across multiple platforms?

### ***Section 5: Final Thoughts & Future Trends***

17. How do you see audience viewership trends evolving in the next five years?
18. What improvements would you like to see in content delivery, engagement, or personalization?
19. Any additional thoughts on how sentiment analysis could help media platforms improve?

### **Conclusion**

Thank you very much for their time and insights. Your response will help in analyzing audience viewership will contribute to the study on sentiment analysis in audience viewership. Finally, would like to receive a summary of the findings once the study is complete? If yes, let us know in the email via [mailchamp@gmail.com](mailto:mailchamp@gmail.com).

## REFERENCES

- Adway, A.M. (2023) ‘The Impact of Television Drama in Understanding History’, *Journal of Al-Tamaddun*, 18(1), pp. 67–78. Available at: <https://doi.org/10.22452/JAT.vol18no1.6>.
- Afrifa, S. *et al.* (2022) ‘Mathematical and Machine Learning Models for Groundwater Level Changes : A Systematic Review and Bibliographic Analysis’, *Future Internet* [Preprint]. Available at: <https://doi.org/https://doi.org/10.3390/fi14090259>.
- Afrifa, S., Zhang, T., *et al.* (2023) ‘Climate change impact assessment on groundwater level changes: A study of hybrid model techniques’, *IET Signal Processing*, 17(April), p. e12227. Available at: <https://doi.org/10.1049/sil2.12227>.
- Afrifa, S., Varadarajan, V., *et al.* (2023) ‘Ensemble Machine Learning Techniques for Accurate and Efficient Detection of Botnet Attacks in Connected Computers’, *MDPI Eng*, pp. 650–664. Available at: <https://doi.org/https://doi.org/10.3390/eng4010039>.
- Al-Hashedi, A. *et al.* (2022) ‘Ensemble Classifiers for Arabic Sentiment Analysis of Social Network (Twitter Data) towards COVID-19-Related Conspiracy Theories’, *Applied Computational Intelligence and Soft Computing*, 2022, pp. 1–10. Available at: <https://doi.org/10.1155/2022/6614730>.
- Alabid, N.N. and Katheeth, Z.D. (2021) ‘Sentiment analysis of Twitter posts related to the COVID-19 vaccines’, 24(3), pp. 1727–1734. Available at: <https://doi.org/10.11591/ijeecs.v24.i3.pp1727-1734>.
- AlBadani, B., Shi, R. and Dong, J. (2022) ‘A Novel Machine Learning Approach for Sentiment Analysis on Twitter Incorporating the Universal Language Model Fine-Tuning and SVM’, *Applied System Innovation*, 5(1), p. 13. Available at: <https://doi.org/10.3390/asi5010013>.
- Alcolea-Díaz, G., Marín-Lladó, C. and Cervi, L. (2022) ‘Expansion of the core business of traditional media companies in Spain through SVOD services’, *Communication and Society*, 35(1), pp. 163–175. Available at: <https://doi.org/10.15581/003.35.1.163-175>.
- Alessandrini, M. *et al.* (2023) ‘A Deep Learning Model for Correlation Analysis between Electroencephalography Signal and Speech Stimuli’, *Sensors*, 23(19), pp. 1–9. Available at: <https://doi.org/10.3390/s23198039>.
- Alkahtani, H. and Aldhyani, T.H.H. (2021) ‘Botnet Attack Detection by Using CNN-LSTM Model for Internet of Things Applications’, *Security and Communication Networks*, 2021. Available at: <https://doi.org/10.1155/2021/3806459>.
- Alqurashi, T. (2022) ‘Stance Analysis of Distance Education in the Kingdom of Saudi Arabia during the COVID-19 Pandemic Using Arabic’.
- Appiahene, P. *et al.* (2023) ‘Application of ensemble models approach in anemia detection using images of the palpable palm’, *Medicine in Novel Technology and Devices*, p. 100269. Available at: <https://doi.org/10.1016/j.medntd.2023.100269>.
- Atiqah, N. *et al.* (2021) ‘Multilingual Sentiment Analysis : A Systematic Literature Review Multilingual Sentiment Analysis : A Systematic Literature Review’, (September). Available at: <https://doi.org/10.47836/pjst.29.1.25>.
- Azeez, N.A. *et al.* (2021) ‘Identification and Detection of Cyberbullying on Facebook Using Machine Learning Algorithms’, *Journal of Cases on Information Technology*,

23(4), pp. 1–21. Available at: <https://doi.org/10.4018/JCIT.296254>.

Baccarne, B., Evens, T. and Schuurman, D. (2013) ‘The Television Struggle : an Assessment of Over-the-Top Television Evolutions in a Cable Dominant Market’, *Communications & strategies*, 92(4), pp. 43–61.

Balakrishnan Gokulakrishnan, Pavalanathan Priyanthan, Thiruchittampalam Ragavan, Nadarajah Prasath, As.P. (2012) ‘Opinion Mining and Sentiment Analysis on a Twitter Data Stream’, p. 229.

Batik, M.V. and Demir, M. (2022) ‘The mediating role of binge-watching in the relationship between type D personality and loneliness’, *Health Psychology Report*, 10(3), pp. 155–167. Available at: <https://doi.org/10.5114/hpr.2021.109550>.

Bharti, S. *et al.* (2021) ‘Cyberbullying detection from tweets using deep learning’, *Kybernetes*, 51(9), pp. 2695–2711. Available at: <https://doi.org/10.1108/K-01-2021-0061>.

Bonta, V., Kumares, N. and Janardhan, N. (2019) ‘A Comprehensive Study on Lexicon Based Approaches for Sentiment Analysis’, *Asian Journal of Computer Science and Technology*, 8(S2), pp. 1–6. Available at: <https://doi.org/10.51983/ajcst-2019.8.s2.2037>.

Borchert, M. and Seifert, R. (2023) ‘Systematic analysis of the pharmacological content of the Tatort (scene of crime) TV series from 2019 to 2021’, *Naunyn-Schmiedeberg's Archives of Pharmacology*, 396(9), pp. 1957–1975. Available at: <https://doi.org/10.1007/s00210-023-02427-3>.

Catelli, R., Pelosi, S. and Esposito, M. (2022) ‘Lexicon-Based vs . Bert-Based Sentiment Analysis : A Comparative Study in Italian’.

Chan-Olmsted, S.M. (2019) ‘A Review of Artificial Intelligence Adoptions in the Media Industry’, *JMM International Journal on Media Management*, 21(3–4), pp. 193–215. Available at: <https://doi.org/10.1080/14241277.2019.1695619>.

Chapola, J.C. *et al.* (2023) ‘Knowledge and perceptions about Dolutegravir and Dolutegravir counselling : a qualitative study among women living with HIV’, pp. 1–10.

Colombini, J. and Duncan, D. (2023) ‘Detached retinas: empathy and the transmedial interstices of RAI fiction’, *Studies in European Cinema*, 20(2), pp. 138–154. Available at: <https://doi.org/10.1080/17411548.2023.2184506>.

Dabla, A. (2004) ‘The Role of Information Technology Policies in Promoting Social and Economic Development: The Case of the State of Andhra Pradesh, India’, *The Electronic Journal of Information Systems in Developing Countries*, 19(1), pp. 1–21. Available at: <https://doi.org/10.1002/j.1681-4835.2004.tb00126.x>.

Dashtipour, K. *et al.* (2016) ‘Multilingual Sentiment Analysis : State of the Art and Independent Comparison of Techniques’, *Cognitive Computation*, 8(4), pp. 757–771. Available at: <https://doi.org/10.1007/s12559-016-9415-7>.

Elankath, S.M. and Ramamirtham, S. (2023) ‘Sentiment analysis of Malayalam tweets using bidirectional encoder representations from transformers: a study’, *Indonesian Journal of Electrical Engineering and Computer Science*, 29(3), pp. 1817–1826. Available at: <https://doi.org/10.11591/ijeecs.v29.i3.pp1817-1826>.

Erkılıç, H. and Erkılıç, S.D. (2022) ‘At the Edge of a New Cognitive Mapping: Ethos Bir Başkadir’, *SERIES: International Journal of TV Serial Narratives*, 8(2), pp. 27–40. Available at: <https://doi.org/10.6092/issn.2421-454X/15401>.

Fägersten, K.B. and Bednarek, M. (2022) ‘The evolution of swearing in television

catchphrases', *Language and Literature*, 31(2), pp. 196–226. Available at: <https://doi.org/10.1177/09639470221090371>.

Faruque, M.A. *et al.* (2021) 'Ascertaining polarity of public opinions on Bangladesh cricket using machine learning techniques', *Spatial Information Research*, 30(1), pp. 1–8. Available at: <https://doi.org/10.1007/s41324-021-00403-8>.

Feng, T., Wang, C. and Chen, C. (2022) 'Creative Effect of Film and Television Advertising Based on Digital Media Interactive Technology', *Mobile Information Systems*, 2022. Available at: <https://doi.org/10.1155/2022/6231760>.

Firoozabadi, A.D. *et al.* (2020) 'Evaluation of localization precision by proposed quasi-spherical nested microphone array in combination with multiresolution adaptive steered response power', *Journal of Electrical Engineering*, 71(3), pp. 150–164. Available at: <https://doi.org/10.2478/jee-2020-0022>.

Gajjar, H. *et al.* (2024) 'A Comparative Analysis of Various Deep-Learning Models for Noise Suppression', *EAI Endorsed Transactions on Internet of Things*, 10, pp. 1–9. Available at: <https://doi.org/10.4108/eetiot.4502>.

Gamal, D. *et al.* (2019) 'Twitter Benchmark Dataset for Arabic Sentiment Analysis', *International Journal of Modern Education and Computer Science*, 11(1), pp. 33–38. Available at: <https://doi.org/10.5815/ijmeecs.2019.01.04>.

Garg, N. and Sharma, K. (2022) 'Text pre-processing of multilingual for sentiment analysis based on social network data', 12(1), pp. 776–784. Available at: <https://doi.org/10.11591/ijece.v12i1.pp776-784>.

Gascón-Vera, P. and Marta-Lazo, C. (2023) 'Formula for the success of humor journalism formats on television according to their professional teams', *Profesional de la Informacion*, 32(2), pp. 1–19. Available at: <https://doi.org/10.3145/EPI.2023.MAR.01>.

Guerini, M., Gatti, L. and Turchi, M. (2013) 'Sentiment analysis: How to derive prior polarities from SentiWordNet', *EMNLP 2013 - 2013 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, (October), pp. 1259–1269.

H. Manguri, K., N. Ramadhan, R. and R. Mohammed Amin, P. (2020) 'Twitter Sentiment Analysis on Worldwide COVID-19 Outbreaks', *Kurdistan Journal of Applied Research*, pp. 54–65. Available at: <https://doi.org/10.24017/covid.8>.

Ham, J. *et al.* (2022) 'Vowel speech recognition from rat electroencephalography using long short-term memory neural network', *PLoS ONE*, 17(6 June), pp. 1–20. Available at: <https://doi.org/10.1371/journal.pone.0270405>.

Harbin, M.B. (2023) 'Don't Make My Entertainment Political! Social Media Responses to Narratives of Racial Duty on Competitive Reality Television Series', *Political Communication*, 40(4), pp. 464–483. Available at: <https://doi.org/10.1080/10584609.2023.2195365>.

Hayawi, K. *et al.* (2023) 'Social media bot detection with deep learning methods: a systematic review', *Neural Computing and Applications*, 35(12), pp. 8903–8918. Available at: <https://doi.org/10.1007/s00521-023-08352-z>.

He, Q. *et al.* (2022) 'Assessment of Bidirectional Relationships between Leisure Sedentary Behaviors and Neuropsychiatric Disorders: A Two-Sample Mendelian Randomization Study', *Genes*, 13(6). Available at:

<https://doi.org/10.3390/genes13060962>.

Hegde, N.P. (2022) 'Employee Sentiment Analysis Towards Remote Work during COVID-19 Using Twitter Data', *International Journal of Intelligent Engineering and Systems*, 15(1), pp. 75–84. Available at: <https://doi.org/10.22266/ijies2022.0228.08>.

Helm, J.E. (2021) 'Distributed Internet voting architecture: A thin client approach to Internet voting', *Journal of Information Technology*, 36(2), pp. 128–153. Available at: <https://doi.org/10.1177/0268396220978983>.

Hoang, M., Bihorac, O.A. and Rouces, J. (2019) 'Aspect-Based Sentiment Analysis using BERT', *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pp. 187–196. Available at: <https://www.aclweb.org/anthology/W19-6120>.

Hossain, M. *et al.* (2025) 'Multi task opinion enhanced hybrid BERT model for mental health analysis', *Scientific Reports*, pp. 1–20. Available at: <https://doi.org/https://doi.org/10.1038/s41598-025-86124-6> 1.

Hu, H. *et al.* (2022) 'A Practical Anonymous Voting Scheme Based on Blockchain for Internet of Energy', *Security and Communication Networks*, 2022(ii). Available at: <https://doi.org/10.1155/2022/4436824>.

Huelin, T. (2022) "'How the Music was Made": Television, musicology and BBC Four', *Critical Studies in Television*, 17(2), pp. 194–200. Available at: <https://doi.org/10.1177/17496020221078503>.

Iordache, C., Raats, T. and Afilipoaie, A. (2022) 'Transnationalisation revisited through the Netflix Original: An analysis of investment strategies in Europe', *Convergence*, 28(1), pp. 236–254. Available at: <https://doi.org/10.1177/13548565211047344>.

Ismond, K.P. *et al.* (2021) 'Assessing Patient Proficiency with Internet-Connected Technology and Their Preferences for E-Health in Cirrhosis', *Journal of Medical Systems*, 45(7). Available at: <https://doi.org/10.1007/s10916-021-01746-3>.

Jin, W. (2020) 'Research on Machine Learning and Its Algorithms and Development', *Journal of Physics: Conference Series*, 1544(1). Available at: <https://doi.org/10.1088/1742-6596/1544/1/012003>.

Jurek, A., Mulvenna, M.D. and Bi, Y. (2015) 'Improved lexicon - based sentiment analysis for social media analytics', *Security Informatics* [Preprint]. Available at: <https://doi.org/10.1186/s13388-015-0024-x>.

Kawade, D.R. and Oza, K.S. (2017) 'Sentiment Analysis : Machine Learning Approach', (June). Available at: <https://doi.org/10.21817/ijet/2017/v9i3/1709030151>.

Killian, G. and McManus, K. (2015) 'A marketing communications approach for the digital era: Managerial guidelines for social media integration', *Business Horizons*, 58(5), pp. 539–549. Available at: <https://doi.org/10.1016/j.bushor.2015.05.006>.

Kim, K. *et al.* (2022) 'Broadcaster Choice and Audience Demand for Live Sport Games: Panel Analyses of the Korea Baseball Organization', *Journal of Sport Management*, 36(5), pp. 488–499. Available at: <https://doi.org/10.1123/jsm.2020-0311>.

Kim, T. (2022) 'Changes and continuities of Makjang drama in the Korean broadcasting industry', *Journal of Japanese and Korean Cinema*, 14(2), pp. 114–130. Available at: <https://doi.org/10.1080/17564905.2022.2124029>.

Kovačević, P. and Perišin, T. (2022) 'Models of TV newsroom organization and news routines in Croatia: Case studies of HRT, Nova TV & N1', *Medijske Studije*, 13(25), pp.

66–89. Available at: <https://doi.org/10.20901/ms.13.25.4>.

Kumar, S., Kumar, M.A. and Soman, K.P. (2019) ‘Deep learning based part-of-speech tagging for Malayalam twitter data (Special issue: Deep learning techniques for natural language processing)’, *Journal of Intelligent Systems*, 28(3), pp. 423–435. Available at: <https://doi.org/10.1515/jisys-2017-0520>.

Kusal, S. *et al.* (2021) ‘AI Based Emotion Detection for Textual Big Data : Techniques and Contribution’, *Big Data and Cognitive Computing* [Preprint]. Available at: <https://doi.org/https://doi.org/10.3390/bdcc5030043>.

Lacasa, P., Martínez-Borda, R. and Lara, I.B. (2022) ‘Transmedia Narratives and Social Networks: Peaky Blinders’ Television Fiction’, *International Journal of Film and Media Arts*, 7(2), pp. 53–73. Available at: <https://doi.org/10.24140/ijfma.v7.n2.03>.

Lengkeek, M., van der Knaap, F. and Frasinca, F. (2023) ‘Leveraging hierarchical language models for aspect-based sentiment analysis on financial data’, *Information Processing & Management*, 60(5), p. 103435. Available at: <https://doi.org/10.1016/j.ipm.2023.103435>.

Li, Y. *et al.* (2023) ‘Enabling Real-Time On-Chip Audio Super Resolution for Bone-Conduction Microphones’, *Sensors*, 23(1), pp. 1–19. Available at: <https://doi.org/10.3390/s23010035>.

Lin, H. *et al.* (2023) ‘PS-Mixer: A Polar-Vector and Strength-Vector Mixer Model for Multimodal Sentiment Analysis’, *Information Processing and Management*, 60(2), p. 103229. Available at: <https://doi.org/10.1016/j.ipm.2022.103229>.

Lye, S.H. and Teh, P.L. (2021) ‘Customer Intent Prediction using Sentiment Analysis Techniques’, pp. 185–190.

Ma, L. and Sun, B. (2020) ‘Machine learning and AI in marketing – Connecting computing power to human insights’, *International Journal of Research in Marketing*, 37(3), pp. 481–504. Available at: <https://doi.org/10.1016/j.ijresmar.2020.04.005>.

Maharani, W. and Effendy, V. (2022) ‘Big five personality prediction based in Indonesian tweets using machine learning methods’, *International Journal of Electrical and Computer Engineering*, 12(2), pp. 1973–1981. Available at: <https://doi.org/10.11591/ijece.v12i2.pp1973-1981>.

Maher, S. and Cake, S. (2023) ‘Innovation in true crime: generic transformation in documentary series’, *Studies in Australasian Cinema*, 17(1–2), pp. 95–109. Available at: <https://doi.org/10.1080/17503175.2023.2224617>.

Massey, P.M. *et al.* (2022) ‘Measuring impact of storyline engagement on health knowledge, attitudes, and norms: A digital evaluation of an online health-focused serial drama in West Africa’, *Journal of Global Health*, 12, pp. 1–10. Available at: <https://doi.org/10.7189/jogh.12.04039>.

Medina, M., Diego, P. and Portilla, I. (2022) ‘Are Video Streaming Platforms Stifling Local Production Creativity? the Spanish Case’, *Creativity*, 9(2), pp. 138–155. Available at: <https://doi.org/10.2478/ctra-2022-0015>.

Miller, K.C. and Nelson, J.L. (2022) ‘“Dark Participation” Without Representation : A Structural Approach to Journalism ’ s Social Media Crisis’. Available at: <https://doi.org/10.1177/20563051221129156>.

Model, C., Ou-yang, C. and Chou, S. (2022) ‘Improving the Forecasting Performance of

Taiwan Car Sales Movement Direction Using Online Sentiment Data and', *Applied Sciences* [Preprint]. Available at: <https://doi.org/https://doi.org/10.3390/app12031550>.

Nauta, A. *et al.* (2022) 'Becoming a Good Chinese Father – Reality TV in China and its Reception', *Global Media and China*, 7(4), pp. 385–399. Available at: <https://doi.org/10.1177/20594364221132150>.

Nguyen, H., Al, R. and Academy, K. (2018) 'Comparative Study of Sentiment Analysis with Product Reviews Using Machine Learning and Lexicon-Based Approaches', 1(4).

Olivieri, M. *et al.* (2024) 'Physics-Informed Neural Network for Volumetric Sound field Reconstruction of Speech Signals', *EURASIP Journal on Audio, Speech, and Music Processing*, 2. Available at: <https://doi.org/10.1186/s13636-024-00366-2>.

Parry, K. and Pitchford-Hyde, J. (2023) "'We may have bad days.. that doesn't make us killers": How military veterans perceive contemporary British media representations of military and post-military life', *Media, War and Conflict*, 16(3), pp. 440–458. Available at: <https://doi.org/10.1177/17506352221113958>.

Pedrero-Esteban, L.-M., Terol-Bolinches, R. and Arense-Gómez, A. (2023) 'El podcast como extensión transmedia sonora de la ficción audiovisual', *Revista Mediterránea de Comunicación*, 14(1), p. 189. Available at: <https://doi.org/10.14198/medcom.23292>.

Ponce, E.K., Cruz, M.F. and Andrade-arenas, L. (2022) 'Machine Learning Applied to Prevention and Mental Health Care in Peru', 13(1).

Rahte, E.Ç. (2022) 'Transnational Audiences of Turkish Dramas: the Case of Sweden', *SERIES: International Journal of TV Serial Narratives*, 8(2), pp. 41–60. Available at: <https://doi.org/10.6092/issn.2421-454X/15488>.

Raj, M. *et al.* (2022) 'An Application to Detect Cyberbullying Using Machine Learning and Deep Learning Techniques', *SN Computer Science*, 3(5), pp. 1–13. Available at: <https://doi.org/10.1007/s42979-022-01308-5>.

Ray, P. and Chakrabarti, A. (2022) 'A Mixed approach of Deep Learning method and Rule-Based method to improve Aspect Level Sentiment Analysis', *Applied Computing and Informatics*, 18(1–2), pp. 163–178. Available at: <https://doi.org/10.1016/j.aci.2019.02.002>.

Ren, Y. *et al.* (2022) 'Label distribution for multimodal machine learning', *Frontiers of Computer Science*, 16(1). Available at: <https://doi.org/10.1007/s11704-021-0611-6>.

Ren, Z. *et al.* (2022) 'Rendered Image Superresolution Reconstruction with Multichannel Feature Network', *Scientific Programming*, 2022. Available at: <https://doi.org/10.1155/2022/9393589>.

Rezaul, K.M. *et al.* (2024) 'Enhancing Audio Classification Through MFCC Feature Extraction and Data Augmentation with CNN and RNN Models', *International Journal of Advanced Computer Science and Applications*, 15(7), pp. 37–53. Available at: <https://doi.org/10.14569/IJACSA.2024.0150704>.

Rintyarna, B.S. *et al.* (2022) 'Modelling Service Quality of Internet Service Providers during COVID-19 : The Customer Perspective Based on Twitter Dataset', pp. 1–12.

Rivas, R. *et al.* (2022) 'Task-agnostic representation learning of multimodal twitter data for downstream applications', *Journal of Big Data*, 9(1). Available at: <https://doi.org/10.1186/s40537-022-00570-x>.

Sadana, M. and Sharma, D. (2020) 'How over-the-top (OTT) platforms engage young

consumers over traditional pay television service? An analysis of changing consumer preferences and gamification', *Young Consumers*, 22(3), pp. 348–367. Available at: <https://doi.org/10.1108/YC-10-2020-1231>.

Schwenk, R.A., Wyss, C. and Aubry, E.M. (2025) 'Experiencing weight stigma during childbirth increases the odds of cesarean birth', *BMC Pregnancy and Childbirth*, 6. Available at: <https://doi.org/https://doi.org/10.1186/s12884-025-07251-6>.

Scott, A.H. and Paprocki, M. (2023) 'Casting Black Athenas: Black Representation of Ancient Greek Goddesses in Modern Audiovisual Media and Beyond', *Journal of Popular Film and Television*, 51(1), pp. 29–38. Available at: <https://doi.org/10.1080/01956051.2023.2171659>.

Shang, L. *et al.* (2023) 'A Lexicon Enhanced Collaborative Network for targeted financial sentiment analysis', *Information Processing and Management*, 60(2), p. 103187. Available at: <https://doi.org/10.1016/j.ipm.2022.103187>.

Sherif, S.M. *et al.* (2023) 'Lexicon annotation in sentiment analysis for dialectal Arabic: Systematic review of current trends and future directions', *Information Processing and Management*, 60(5), p. 103449. Available at: <https://doi.org/10.1016/j.ipm.2023.103449>.

Shiva, L. *et al.* (2021) 'Negative Childbirth Experience and Post-traumatic Stress Disorder - A Study Among Postpartum Women in South India', *Frontiers in Psychiatry*, 12(July), pp. 1–7. Available at: <https://doi.org/10.3389/fpsy.2021.640014>.

Singh, A. and Glińska-Neweś, A. (2022) 'Modeling the public attitude towards organic foods: a big data and text mining approach', *Journal of Big Data*, 9(1). Available at: <https://doi.org/10.1186/s40537-021-00551-6>.

Singh, P., Sawhney, R.S. and Kahlon, K.S. (2018) 'Sentiment analysis of demonetization of 500 & 1000 rupee banknotes by Indian government', *ICT Express*, 4(3), pp. 124–129. Available at: <https://doi.org/10.1016/j.icte.2017.03.001>.

Suganthi, D. (2024) 'Predicting Postpartum Depression with Aid of Social Media Texts Using Optimized Machine Learning Model', 17(3), pp. 417–427. Available at: <https://doi.org/10.22266/ijies2024.0630.33>.

Taboada, M., Brooke, J. and Voll, K. (2022) 'Lexicon-Based Methods for Sentiment Analysis', (September 2010).

Tâm, T. *et al.* (2016) *Television Audiences Across the World*.

Tanriöver, H.U. (2022) 'Towards a Social History of Turkey Through Television Series', *SERIES: International Journal of TV Serial Narratives*, 8(2), pp. 9–26. Available at: <https://doi.org/10.6092/issn.2421-454X/15676>.

Uddin, M.N. and Hafiz, F. Bin (2022) 'Drug Sentiment Analysis using Machine Learning Classifiers', 13(1), pp. 92–100.

Valtorta, R.R. *et al.* (2023) 'Gender Stereotypes and Sexualization in Italian Children's Television Advertisements', *Sexuality and Culture*, 27(5), pp. 1625–1645. Available at: <https://doi.org/10.1007/s12119-023-10081-3>.

Verma, R., Chhabra, A. and Gupta, A. (2023) 'A statistical analysis of tweets on covid-19 vaccine hesitancy utilizing opinion mining: an Indian perspective', *Social Network Analysis and Mining*, 13(1), pp. 1–12. Available at: <https://doi.org/10.1007/s13278-022-01015-2>.

Vodičková, K. (2022) 'Impact of Global Streaming Platforms on Television Production:

A Case Study of Czech Content Production', *Medijske Studije*, 13(26), pp. 27–47. Available at: <https://doi.org/10.20901/ms.13.26.2>.

Wang, Z. *et al.* (2017) 'Fine-grained sentiment analysis of social media with emotion sensing', *FTC 2016 - Proceedings of Future Technologies Conference*, (December), pp. 1361–1364. Available at: <https://doi.org/10.1109/FTC.2016.7821783>.

Wayne, M.L. (2022) 'Netflix audience data, streaming industry discourse, and the emerging realities of “popular” television', *Media, Culture and Society*, 44(2), pp. 193–209. Available at: <https://doi.org/10.1177/01634437211022723>.

Wayne, M.L. and Castro, D. (2021) 'SVOD Global Expansion in Cross-National Comparative Perspective: Netflix in Israel and Spain', *Television and New Media*, 22(8), pp. 896–913. Available at: <https://doi.org/10.1177/1527476420926496>.

Wunderlich, F. and Memmert, D. (2022) 'A big data analysis of Twitter data during premier league matches: do tweets contain information valuable for in-play forecasting of goals in football?', *Social Network Analysis and Mining*, 12(1), pp. 1–15. Available at: <https://doi.org/10.1007/s13278-021-00842-z>.

Xiang, M. *et al.* (2023) 'A Study on Online Health Community Users ' Information Demands Based on the BERT-LDA Model', *Healthcare (Switzerland)* [Preprint]. Available at: <https://doi.org/https://doi.org/10.3390/healthcare11152142>.

Xu, Q. *et al.* (2022) 'A Dual-Pointer guided transition system for end-to-end structured sentiment analysis with global graph reasoning', *Information Processing and Management*, 59(4), p. 102992. Available at: <https://doi.org/10.1016/j.ipm.2022.102992>.

Yadav, N.N. and D. (2018) 'Lexicon-based approach to Sentiment Analysis of tweets using R language', (April). Available at: <https://doi.org/10.1007/978-981-13-1810-8>.

Zhang, H., Gan, W. and Jiang, B. (2014) 'Machine Learning and Lexicon based Methods for Sentiment Classification : A Survey 1', pp. 262–265. Available at: <https://doi.org/10.1109/WISA.2014.55>.