LUISS Guido Carli
Department of Business and Management

Bachelor's Degree in Management & Computer Science

The Influence of Online Community Dynamics on Brand and Creator Strategy

Final Thesis by: Lorenzo Laterza
Student ID: 283121
Supervisor: Prof. Irene Finocchi

Academic Year 2024/2025

Acknowledgements

I would like to express my sincere gratitude to several individuals who provided invaluable guidance and support in completing this thesis.

First and foremost, I would like to thank my supervisor, Professor Irene Finocchi. Her constant enthusiasm for the project was a significant source of encouragement at every stage. I am particularly grateful for the intellectual freedom and trust she gave me, which allowed this research to evolve organically while she always provided valuable support. Her guidance fostered an ideal academic environment in which to develop this work.

I would also like to express my deepest gratitude to the entire team at Viralba for welcoming this project and providing an invaluable real-world context for my research. Special thanks are reserved for Alessandro Chessa. Not only did he grant me the great opportunity to work with him on this project, he also showed a genuine interest and an immediate willingness to collaborate from our very first meeting — a quality that is as rare as it is valuable and should never be taken for granted. The trust and consideration he has consistently shown in my abilities has been a constant source of motivation. His guidance was invaluable, helping me to find the 'story' within the numbers and constantly bridging the gap between rigorous academic research and real-world business application.

I am also immensely grateful to Antonio Agabio. His insightful advice on the thesis's narrative structure was invaluable. He helped me see the bigger picture, transforming what could have been a series of disconnected analyses into a single, coherent and compelling argument.

The mentorship I received from Alessandro and Antonio extended far beyond the pages of this thesis, providing invaluable professional and personal guidance for my future career. I sincerely hope that this is the first of many collaborations to come.

Abstract

The growing influence of online communities presents a significant strategic challenge for brands and creators. This thesis addresses the inadequacy of traditional analytics, which often rely on superficial "vanity metrics" and fail to capture the complex dynamics of audience behavior. To overcome this, the research proposes and validates an integrated analytical framework. The methodology combines two complementary approaches: first, an AI-powered content analysis to understand what a community discusses and how it feels, utilizing Sentiment, Emotional, and advanced Topic Detection models; and second, Social Network Analysis to understand who drives the conversation and how influence propagates. This dual framework was applied to a large-scale case study of the sports streaming service DAZN, examining over 22,000 social media posts to map its community structure. The findings demonstrate the framework's efficacy in generating deep, non-obvious insights. The analysis successfully identified and categorized a cohort of previously hidden, independent micro-influencers into distinct functional archetypes (e.g., Community Leaders, Information Brokers). Furthermore, a comparative analysis of a brand-related controversy revealed a significant 'power shift' in the influence hierarchy, demonstrating that the community gravitates towards specialist authorities in times of debate, a dynamic invisible to standard metrics. The thesis concludes that this integrated approach provides a robust method for decoding the influence of community dynamics. It offers a clear pathway for brands and creators to move beyond passive measurement towards a proactive, data-informed strategic process, enabling them to understand and engage with their audiences with unprecedented clarity and precision.

Contents

Acknowledgements **1**

Abstract **2**

Introduction **5**

1 Understanding the Social Media Landscape and the Limits of Traditional Metrics **7**
1.1 The Modern Conversational Landscape in a Pervasive Ecosystem 7
1.2 Key Actors in the Social Media Ecosystem 8
1.2.1 Different Objectives and Success Metrics 8
1.2.2 Communication Styles and Content Strategies 9
1.2.3 Audience Reach and Targeting Approaches 10
1.2.4 Trust, Credibility and Influencer Power 10
1.2.5 The Empowered Consumer in a Three-Way Ecosystem 11
1.3 The "Vanity Metrics Trap" and the Limits of Traditional Analytics 12
1.4 The Need for a Paradigm Shift from Metrics to Intelligence 13

2 From Language Models to Insights, The Viralba Approach **15**
2.1 Viralba's Vision and the Underlying Technology 15
2.2 Understanding Modern NLP Foundations 16
2.2.1 The Impact of the Transformer Architecture 16
2.2.2 BERT and the Evolution of Pretrained Models 17
2.3 Measuring Tone in Social Media Conversation 18
2.3.1 Fine-Tuning for Sentiment and Emotion Analysis 18
2.3.2 Adapting XLM-T for Viralba's Sentiment Detection 19
2.3.3 Capturing Emotions Behind User Reactions 20
2.4 Detecting Conversation Topics with BERTopic 20
2.4.1 Data Cleaning and Preparation for Topic Modeling 20
2.4.2 From Sentence Embeddings to Topic Clusters 21
2.4.3 How Viralba Optimizes and Labels Topics with AI 22
2.5 Bringing Sentiment and Topics Together 23

3 Testing the Potential of Network Analysis for Brands **24**
3.1 Extending Viralba with Network Analysis 24
3.2 Collecting and Structuring the Dataset for Social Graphs 24
3.3 Analyzing the User Network Around DAZN 27
3.3.1 Algorithms for Calculating Influence and Detecting Communities . 28

3.3.2 Institutional and Official Accounts at the Core of the Network . . . 29
3.3.3 Finding Independent Voices Through Strategic Filtering 29
3.3.4 Understanding the Roles of Key Users in the Network 30
3.4 Exploring Topics Through Hashtags 31
3.5 Focusing on the Subscription Debate 34
3.5.1 Structural Shifts and the Rise of Specialist Accounts 35
3.6 Future Applications of Network Analysis in Viralba's Strategy 37

Conclusion **38**

References **40**

Introduction

The 21st century has witnessed a fundamental re-architecting of the relationship between organizations, public figures, and their audiences. The rise of a vast and interconnected digital ecosystem, powered by social media platforms, has shifted the center of gravity of influence. Online communities are no longer passive recipients of broadcasted messages; they are active participants in the forging of public opinion, the co-creation of brand narratives, and the rise and fall of creator reputations. For both established brands and independent creators, understanding and navigating this new landscape is no longer a peripheral marketing activity, but a central strategic imperative. The influence of these online community dynamics is undeniable, yet the methods for decoding this influence remain a significant challenge.

Despite this recognized importance, the primary tools used to measure success in this new environment often prove to be inadequate. Brands and creators alike have become reliant on a suite of traditional analytics—often referred to as "vanity metrics"—such as follower counts, likes, and reach. While easy to measure, these indicators primarily quantify the volume of online activity, offering little insight into the quality or substance of community engagement. They fail to answer the most critical strategic questions: What is the underlying sentiment of the conversation? What are the key topics driving community interest or concern? And who are the true opinion leaders shaping the discourse? Relying on these superficial metrics is akin to navigating a complex social landscape with an incomplete map.

To move beyond these limitations, this thesis argues that a more sophisticated analytical approach is required—one that can provide deep, actionable intelligence. It posits that a true understanding of online community dynamics necessitates a dual framework that analyzes both the *content* and the *structure* of conversations. First, it requires an understanding of *what* is being said and *how* it is being said, a challenge addressed by advanced Artificial Intelligence algorithms for Sentiment, Emotional, and Topic Analysis. This initial layer of analysis allows for a direct measurement of a community's affective responses and thematic interests, providing a rich, qualitative understanding that raw numbers cannot capture.

However, understanding the content alone is insufficient. It is equally crucial to understand *who* is driving the conversation and *how* their ideas propagate, a challenge addressed by Social Network Analysis (SNA). This second analytical layer moves beyond the text to map the relational architecture of the community, identifying key influencers, bridges between groups, and the overall flow of information. The true power lies not in using these methodologies in isolation, but in their strategic integration.

This thesis therefore seeks to answer the following central research question: *How can*

an analytical framework that combines AI-based content analysis with Social Network Analysis be used to decode the influence of online community dynamics, thereby providing actionable intelligence to inform brand and creator strategy?

To address this question, the research is structured into three main chapters. Chapter 1 will establish the context by analyzing the social media ecosystem, its key actors—brands and creators—and the critical limitations of traditional analytics. Chapter 2 will then provide a deep dive into the AI algorithms that form the core of a modern content intelligence platform, explaining the technological foundations from the Transformer architecture to its specific applications. Finally, Chapter 3 will apply Social Network Analysis in a novel case study of a major brand, demonstrating how this integrated framework can be used to map the architecture of a real-world community, identify its most influential members, and uncover the nuanced, context-dependent nature of influence itself.

Chapter 1

Understanding the Social Media Landscape and the Limits of Traditional Metrics

1.1 The Modern Conversational Landscape in a Pervasive Ecosystem

The dawn of the 21st century has been marked by a profound transformation in how humanity communicates, connects, and consumes information. Central to this shift is the rise of a pervasive and dynamic digital ecosystem commonly referred to as social media. Its scale is difficult to overstate: as of early 2025, there are over 5.24 billion active social media user identities worldwide, a figure equivalent to nearly 64% of the total global population. The average user now spends 2 hours and 21 minutes per day on social platforms, making it a dominant force in the daily allocation of human attention Kemp, 2025. This is not merely a technological trend, but a fundamental re-architecting of social and commercial interaction, creating a new landscape that presents unprecedented opportunities and challenges for the individuals, creators, and brands that inhabit it.

To understand this landscape, it is useful to deconstruct its core components. Kietzmann et al., 2011 proposed a foundational framework describing social media through seven functional "building blocks": identity, conversations, sharing, presence, relationships, reputation, and groups. These blocks represent the different facets of user experience and platform functionality. Different social media sites are not monolithic entities, but rather unique combinations of these blocks, each emphasizing certain functions over others. For example, a professional network like LinkedIn heavily prioritizes identity, relationships, and reputation, while a platform like Instagram focuses more on sharing and presence. This framework highlights that "social media" is not a single entity, but a complex interplay of functionalities that users can leverage to construct their digital presence and interact with communities.

Historically, the most disruptive feature of this new paradigm was its ability to facilitate the creation and exchange of User-Generated Content (UGC), fundamentally altering the flow of information Kietzmann et al., 2011. This marked a critical departure from the one-to-many broadcast model of traditional media. Power shifted from institutions to

individuals; consumers, who had historically been passive recipients of content, were now empowered to create, modify, and discuss it. The infamous "United Breaks Guitars" incident of 2009, where a musician's complaint video against an airline went viral on YouTube, became a textbook case of this new "democratized" communication, illustrating how a single creative consumer could inflict significant reputational damage on a multinational corporation Kietzmann et al., 2011.

Over a decade later, this ecosystem has evolved far beyond a simple collection of platforms for UGC. Social media is now more accurately understood as a "technology-centric ecosystem" where a diverse and complex set of behaviors and exchanges occur between interconnected actors Appel et al., 2020. The distinction between "online" and "offline" life has become increasingly porous. Social media functionality is no longer confined to its native applications but is deeply embedded into nearly every facet of digital existence, from logging into a news site with a Facebook account to sharing a song from Spotify directly to an Instagram story. This has created what Appel et al., 2020 term an "omni-social" world, where consumer decision-making, from need recognition to post-purchase evaluation, is perpetually influenced by social touchpoints.

A defining feature of this new, omni-social landscape is the reconfiguration of who holds authority and trust. Alongside the brands themselves, which must navigate this complex environment, a new class of powerful digital actors has become firmly established: the social media creator, or influencer. This phenomenon has evolved the traditional celebrity endorsement into a more pervasive and nuanced form of communication. The most significant development within this trend is the rise of the "micro-influencer." Unlike traditional celebrities with massive, broad followings, micro-influencers are creators with smaller, more targeted audiences. Their power does not lie in sheer reach, but in their perceived credibility and authenticity. They are often seen as trusted "experts" in their niche, and their first-person narration is considered warmer and more personal than traditional advertising, making them highly effective at engaging consumers Appel et al., 2020. The concept of the micro-influencer, as we will demonstrate in the analytical chapter of this thesis, is not merely a qualitative observation but a structurally identifiable role that can be uncovered through network analysis. Understanding the distinct objectives and strategies of these two key actors, the centrally broadcasting brand and the community-embedded creator, is therefore the essential next step in dissecting the modern social media ecosystem.

1.2 Key Actors in the Social Media Ecosystem

Within the social media ecosystem, the official profiles of brands and those of creators, or influencers, play distinct yet complementary roles. Both operate to engage consumers and shape perceptions, but they do so with fundamentally different objectives, communicative strategies, and measures of success. Understanding this distinction is the first step toward dissecting the complex dynamics of modern digital marketing.

1.2.1 Different Objectives and Success Metrics

For corporate entities, social media is a strategic tool integrated into broader business objectives. Brands utilize these platforms primarily to increase brand awareness, manage

their reputation, stimulate consumer engagement, and, ultimately, drive sales and revenue. The success of these efforts is not left to chance but is rigorously measured through a set of quantitative key performance indicators (KPIs). As will be discussed in the following section, these metrics range from gauging visibility, such as reach and impressions, to performance-oriented indicators like conversion rates and, most critically, return on investment (ROI) Kočišová and Štarchoň, 2023. The brand's perspective is inherently tied to scalable and quantifiable business outcomes.

In contrast, the objectives of creators are intrinsically linked to the development of their personal brand and the cultivation of a loyal community. Their primary goal is to grow their personal audience and maintain a strong, authentic bond with their followers. This personal brand then becomes a monetizable asset through two primary avenues. Firstly, creators with sufficient influence may be approached by companies to act as brand ambassadors or to create sponsored content, for which they are compensated. Secondly, the most successful creators can leverage their credibility to launch their own product lines, effectively transforming themselves into standalone brands Schaffer, 2025. Consequently, success for a creator is measured by metrics that reflect community health and persuasive power, such as the growth of their follower count, their engagement rate, and their demonstrated ability to influence audience behavior. This highlights a fundamental strategic divergence: while brands often pursue scale and broad market penetration, creators leverage the power of niche communities and the invaluable currency of personal credibility Landingi, 2025.

1.2.2 Communication Styles and Content Strategies

These divergent objectives naturally manifest in distinct communicative styles and content strategies. The way a brand communicates is fundamentally different from the way a creator interacts with their audience, a distinction that lies at the heart of their respective roles in the social media ecosystem.

Brands, in their capacity as official corporate entities, typically produce content that is carefully planned, polished, and consistent with established communication guidelines. Their output is often informational, promotional, or institutional in nature, designed to convey a message of authority, reliability, and quality. The tone is generally controlled, and the primary goal is to manage the brand's image and present its products or services in the best possible light. This approach, while professional, often creates a perceptible distance between the company and the consumer.

Creators, in stark contrast, operate on a foundation of personal narrative and perceived authenticity. Their communicative style is more informal, spontaneous, and conversational, aiming to build a relatable persona. Instead of creating formal advertisements, they integrate products and brands into the context of their own daily lives and personal experiences, a method that is perceived by audiences as more genuine Barquero Cabrero et al., 2023. A creator does not simply present a product; they tell a story about it, framing it as a personal discovery or a trusted recommendation. This narrative approach is key to building the strong, personal connection they have with their followers.

This fundamental difference in authenticity has a direct impact on how content is received by audiences. While brand-generated content is often recognized and processed as advertising, creator content can feel more akin to a personal endorsement from a trusted

peer. This perception is corroborated by recent industry data, which indicates that a significant portion of companies, 36% in a recent survey, now consider content produced by influencers to be more effective than their own brand-created assets Matter Communications, 2023. The reason is clear: an authentic, personal narrative can bypass the natural skepticism that consumers have towards direct advertising, making the message more persuasive and memorable.

1.2.3 Audience Reach and Targeting Approaches

The strategic differences between brands and creators are further reflected in how they approach audience reach and targeting. While both seek to deliver their message to a relevant audience, their methodologies and the nature of their reach are fundamentally distinct.

Brands typically pursue a strategy of broad, scalable reach. Leveraging the sophisticated advertising tools offered by social media platforms, they aim to connect with large, diversified market segments. Their targeting is often based on demographic data (age, location, gender), psychographic profiles (interests, behaviors), or past interactions with the brand. The primary goal of this approach is to maximize visibility across a wide and often generalist audience, ensuring that the brand message is seen by the largest possible number of potential customers.

Creators, in contrast, operate within more circumscribed and vertical communities. Their audience is not built through broad advertising campaigns but is cultivated organically around a shared interest, passion, or identity. A creator's reach is therefore narrower but significantly deeper. They are not broadcasting to a market segment; they are communicating within a niche community. This structural difference makes creators exceptionally valuable for highly targeted campaigns designed to reach audiences that are often difficult to access through traditional corporate channels Landingi, 2025. For a brand, collaborating with a creator is a way to "borrow" their deep, trusted connection with a specific, highly-engaged community, thereby delivering a message with a level of precision and resonance that broad-based advertising struggles to achieve.

1.2.4 Trust, Credibility and Influencer Power

Perhaps the most critical dimension differentiating brands from creators is the perceived trust that they command. In an information-saturated environment, consumer skepticism towards direct advertising is a significant hurdle for brands to overcome. The messages they broadcast, being inherently commercial in nature, are often met with a degree of distrust. While brands work to build credibility through consistency, quality, and customer service, their communication is fundamentally understood by consumers as a form of advertising.

Creators, on the other hand, operate with a different currency: authenticity. They cultivate parasocial relationships with their followers, a psychological phenomenon where the audience develops a one-sided, perceived friendship with a media figure. This bond is not built on corporate messaging but on a sense of perceived intimacy, shared values, and personal connection. The creator is not seen as an institution, but as a trusted peer, a knowledgeable friend, or an inspirational figure.

This dynamic fundamentally alters how their recommendations are received. While a brand's message is processed as an advertisement, an influencer's endorsement is often perceived as a genuine, personal recommendation. This is strongly supported by consumer data, with one major industry report indicating that approximately 69% of consumers trust recommendations from influencers more than direct communication from a brand Influencer Marketing Hub, 2022. Empirical research further substantiates this, demonstrating that key influencer attributes such as trustworthiness, expertise, and authenticity can significantly enhance a brand's credibility and positively impact consumer purchase intentions Liu and Zheng, 2024. This ability to leverage a trusted, parasocial bond to vouch for a product is the core of an influencer's power and value within the marketing ecosystem.

1.2.5 The Empowered Consumer in a Three-Way Ecosystem

The social media ecosystem is not merely a dyad of brands and creators; it is a triad, with the consumer functioning as a third, and increasingly powerful, key actor. The modern consumer is no longer a passive recipient of messages but an active participant in the online discourse, co-creating content and shaping brand narratives.

Brands actively encourage this participation through various engagement tactics, such as contests, polls, and dedicated customer care channels. However, their communication often remains within an "institutional" frame, where the interaction is clearly delineated as being between a company and its customer. While valuable, this form of engagement is often perceived as structured and managed.

Creators, conversely, tend to generate a more intense and organic form of bidirectional dialogue. Their personal and narrative-driven content naturally stimulates a higher volume of comments, questions, and shares, fostering a sense of genuine conversation within the community. Recent research on Instagram, for instance, empirically demonstrates that user-generated posts receive significantly more comments than corporate content, even when the number of likes is comparable Barquero Cabrero et al., 2023. This finding highlights a crucial insight: consumers are more inclined to interact with content that they perceive as authentic and spontaneous.

This dynamic gives rise to the powerful phenomenon of User-Generated Content (UGC) in relation to a brand. When consumers create their own content featuring a product, whether it's an unboxing video, a product review, or a photo shared on Instagram, they are, in effect, acting as unpaid, highly credible micro-influencers. This form of earned media is often perceived as more trustworthy than either official brand content or sponsored creator content, as it is seen as an unbiased endorsement from a genuine peer. Harnessing and amplifying positive UGC has therefore become a cornerstone of modern community marketing strategies.

In summary, brands, creators, and consumers form a complex and symbiotic ecosystem where influence is fluid and value is co-created. Understanding the distinct roles and motivations of these three actors is the first step toward strategic mastery of the social media landscape. However, simply acknowledging their existence is not enough. The next critical question is: how is success measured in this intricate new environment? While brands and creators alike track a host of metrics to gauge their performance, a closer examination reveals that many of these conventional measures are often superficial,

failing to capture the true health and persuasive power of a community. This leads to the fundamental problem that will be explored in the following section: the trap of "vanity metrics."

1.3 The "Vanity Metrics Trap" and the Limits of Traditional Analytics

The complex, symbiotic ecosystem of brands, creators, and consumers requires a framework for measuring success. In response, a set of conventional metrics has been widely adopted across the industry, forming the backbone of most standard social media analytics dashboards. These metrics are typically grouped into two main categories: those that measure visibility and those that measure interaction.

The first category, focused on visibility, aims to quantify the sheer scale of a message's dissemination. The most common metrics here are Reach, which represents the total number of unique users who have seen a piece of content, and Impressions, the total number of times that content was displayed, regardless of whether it was clicked or not. These indicators are often used as a primary measure of brand awareness Kočišová and Štarchoň, 2023.

The second category focuses on engagement, attempting to measure how actively the audience interacts with the content. This is typically calculated as the Engagement Rate, a percentage derived from the sum of all interactions (such as likes, comments, and shares) divided by the number of followers or the total reach. A high engagement rate is often interpreted as a sign of a healthy, active community and resonant content. These metrics, alongside performance indicators like Click-Through Rate (CTR) and Conversion Rate, have long been the standard for evaluating social media marketing performance.

While these metrics provide a seemingly straightforward dashboard of performance, a growing body of academic and industry critique has highlighted their significant limitations. Many of these conventional indicators fall into the category of "vanity metrics": numbers that are easy to measure and appear impressive on the surface, but which often fail to correlate with, or contribute to, meaningful business outcomes HubSpot, 2025; Rogers, 2018. High reach, for instance, does not guarantee that the content was actually attended to or understood, and a high number of "likes" does not necessarily translate into increased brand loyalty or purchase intent. These metrics measure the *quantity* of exposure but say very little about the *quality* of the engagement.

This is not a new problem. As early as 2010, researchers were already questioning the direct applicability of traditional Return on Investment (ROI) models to the fluid and conversational environment of social media. Hoffman and Fodor, 2010 argued that instead of focusing on the firm's investment, marketers should focus on the consumers' investments of their own time and engagement. They proposed a shift away from simply measuring eyeballs and towards understanding active user contributions, such as comments, content sharing, and other forms of meaningful interaction. This early critique highlighted a fundamental flaw that persists to this day: the standard analytics offered by social platforms are adept at counting actions, but they are ill-equipped to measure deeper, more strategic concepts such as community health, brand trust, or the true persuasive impact of a message.

More than a decade later, this fundamental disconnect between measurement and strategic value has only intensified. Modern review papers in the field of marketing confirm that many current approaches to performance measurement remain insufficient, overly quantitative, and strategically disconnected from real business outcomes Ascani and Ancillai, 2025. The analytics provided by social platforms are designed to measure what is easy to count, not necessarily what is important to understand. They can quantify the size of an audience, but not its trust. They can track the engagement rate on a post, but not the underlying sentiment or the emotional drivers behind that engagement.

This "vanity metrics trap" creates a significant risk for brands and creators: it can lead to the optimization of strategies that are superficially successful but strategically empty. For example, a content strategy might be geared towards maximizing "likes" by posting memes or viral content, succeeding on paper while failing to build any meaningful brand equity or drive long-term customer loyalty. The reliance on these simple, quantitative indicators fails to capture the complex, qualitative dynamics of a a community.

Ultimately, these classic metrics are unable to answer the most critical strategic questions: Who are the true opinion leaders within a community? What are the core topics that genuinely resonate with an audience, and what is their emotional response to them? How do opinions and ideas propagate through the network? Answering these questions requires moving beyond the simple act of counting and towards a deeper, more sophisticated mode of analysis. This necessitates a paradigm shift from a reliance on platform-provided dashboards to the adoption of advanced analytical tools, powered by Artificial Intelligence, capable of interpreting the rich, unstructured data where the true insights lie.

1.4 The Need for a Paradigm Shift from Metrics to Intelligence

The inherent limitations of traditional, volume-based metrics create a significant strategic void. While capable of measuring the scale of online activity, they fail to provide a deep understanding of community health, audience sentiment, or true business impact, leading to what is often described as the "vanity metrics trap" Rogers, 2018. This gap has prompted a necessary and fundamental paradigm shift in the field, moving away from simple social media measurement and towards the more sophisticated discipline of Social Media Intelligence (SMI).

This concept, first established in academic literature over a decade ago, refers to the systematic process of gathering, analysing, and interpreting social media data to generate actionable insights that can inform strategic business decisions Zeng et al., 2010. The impetus for the necessity of SMI is predicated on the sheer magnitude and intricacy of contemporary user-generated data. In the contemporary landscape, brands and creators are no longer engaged with manageable focus groups; rather, they are confronted with a continuous, global stream of unstructured text, images, and videos. The "Big Data" nature of this environment, characterised by immense volume, high velocity, and significant variety, renders manual or simplistic analysis practically impossible. The efficacy of advanced tools in the context of complexity is no longer a matter of speculation; they have been proven to be the most effective means of comprehension.

In response to this challenge, the discipline has evolved into the broader field of Social Me-

dia Analytics. This field is dedicated to developing and applying advanced computational methods and information systems to address the key challenges of processing large-scale social data. Its primary concerns include the automatic discovery of emergent topics, the real-time collection of relevant data, and the sophisticated preparation of that data for analysis Stieglitz et al., 2018. A central practice within this new paradigm is social listening. Unlike passive monitoring, social listening is a proactive approach that aims to understand the context, sentiment, and nuanced drivers behind online conversations. It represents a strategic effort to move beyond simply tracking brand mentions and towards a genuine comprehension of customer needs, pain points, and perceptions, ultimately improving customer engagement and retention in ways that raw metrics cannot Garhwal and Dhanawade, 2023.

The engine that powers this entire shift from passive metrics to active intelligence is Artificial Intelligence. The integration of AI, particularly machine learning and Natural Language Processing (NLP), into marketing workflows is what enables the transformation of chaotic, unstructured data into structured, strategic insights JETIR, 2024. Instead of focusing on likes and shares, modern analytical frameworks, such as those recommended by industry leaders like Sprout Social Sprout Social, 2024, emphasize a new class of evolved, qualitative metrics. These include sentiment scores (the overall positive or negative tone of a conversation), share of voice (a brand's visibility on a topic compared to its competitors), and influence scores (a measure of an individual's ability to drive conversations). These sophisticated indicators cannot be derived from simple counting; they require advanced algorithms capable of interpreting the meaning and context of human language.

Ultimately, this technological and strategic evolution is a direct response to a clear business imperative: the need to connect marketing activities directly to tangible business objectives. As industry reports now emphasize, the true value of analytics lies in its ability to link the KPIs from social media to the overarching goals of the organization, moving beyond vanity metrics to understand true performance and predict future outcomes Deloitte Digital, 2022.

Having thus established the context of the modern social media ecosystem, the key actors that inhabit it, and the critical shortcomings of traditional analytics, the strategic necessity for a new generation of analytical tools becomes clear. The shift from simple metrics to deep intelligence is not merely a theoretical preference but a practical requirement for any brand or creator seeking to compete and thrive in the contemporary digital landscape. The following chapter will therefore provide a comprehensive methodological deep dive into the specific algorithms that form the core of a modern intelligence platform. We will dissect the technological foundations that enable a true understanding of online conversations and explore how a platform like Viralba leverages this technology to deliver the very insights that traditional metrics fail to capture.

Chapter 2

From Language Models to Insights, The Viralba Approach

2.1 Viralba's Vision and the Underlying Technology

As established in the previous chapter, the contemporary social media ecosystem presents a dual challenge for brands and content creators: the traditional metrics for measuring success have proven insufficient, while the sheer volume of data makes manual analysis untenable. This creates a critical need for a new generation of analytical tools capable of moving beyond superficial metrics to provide deep, actionable, and quantitative intelligence. It is precisely to address this need that Viralba, the startup framing the context of this thesis, is developing its technological platform.

Viralba is an Artificial Intelligence-based platform designed to empower content creators and social media agencies. Its core mission is twofold: first, to enable the creation of content that is deeply aligned with the authentic desires of an audience; and second, to provide a sophisticated mechanism for gathering both qualitative and quantitative feedback on how that content is received. The platform operates by collecting and analyzing information in real-time from a spectrum of key social networks, including YouTube, Instagram, Facebook, and X (formerly Twitter).

The fundamental "pain point" that Viralba addresses is the immense challenge creators and agencies face in keeping pace with the ever-evolving online landscape. Understanding what truly attracts an audience, what a community values, and what it rejects requires a continuous process of data gathering and interpretation that can be prohibitively time-consuming. Viralba is therefore engineered to be more than a mere data analysis tool; it is a comprehensive intelligence platform that automates the laborious processes of data collection and computation. By doing so, it allows its users, creators and strategists, to spend less time on computational tasks and more time focusing on their core strengths: creating high-quality content and engaging with their communities.

The platform provides its clients with a direct competitive advantage by enabling data-driven interventions that have a tangible impact on marketing and content strategy. The strategic power of this approach can be illustrated with a practical example. Consider a brand in the sustainable fashion industry. A conventional analysis might suggest collaborating with influencers who explicitly discuss ethical clothing. However, a deep analysis

of the brand's customer community, as performed by Viralba, could reveal an unexpected but strong thematic overlap: a significant portion of their audience also actively participates in discussions around minimalist home design and zero-waste lifestyles. This non-intuitive connection, between sustainable fashion and home minimalism, represents a powerful, untapped strategic insight. It suggests that the underlying value shared by the community is not just "ethical clothing," but a broader principle of "conscious consumerism." Armed with this knowledge, the brand could refine its messaging or identify a whole new category of lifestyle influencers for collaboration, an opportunity that would have been missed by a traditional, category-based approach.

To achieve this vision, Viralba's platform is built upon a sophisticated stack of Artificial Intelligence algorithms. This chapter will dissect this technological core. We will begin by examining the foundational technology that enables a deep understanding of language, the Transformer architecture and the BERT model. We will then explore how Viralba applies this technology to two key analytical tasks: first, to gauge the tone and emotion of conversations through Sentiment and Emotional Analysis, and second, to identify the core themes of discussion using the state-of-the-art BERTopic model.

2.2 Understanding Modern NLP Foundations

The analytical capabilities of a modern intelligence platform like Viralba are built upon a recent but revolutionary technological foundation. To fully appreciate why certain models are so effective, and why they represent a paradigm shift from the tools traditionally offered by social media platforms, it is essential to first understand the architectural innovation that made them possible: the Transformer.

2.2.1 The Impact of the Transformer Architecture

Prior to 2017, the state-of-the-art in Natural Language Processing was dominated by models with a sequential design, most notably Recurrent Neural Networks (RNNs). These models processed text in a way that mimics human reading: one word at a time, from left to right. While intuitive, this sequential nature imposed severe limitations. It created a "memory bottleneck," making it difficult for the models to capture complex relationships between words that were far apart in a long sentence. Furthermore, it fundamentally prevented the parallel processing of words within a single text, creating a significant computational handicap that slowed down training on the massive datasets required to learn the complexities of human language.

This paradigm was shattered by the introduction of the Transformer architecture in a seminal paper by Vaswani et al., 2017, aptly titled "Attention Is All You Need." The Transformer dispensed with recurrence entirely. Its groundbreaking innovation was a mechanism called self-attention, which allows the model to look at all the words in a sentence simultaneously and to weigh their relevance to each other. Instead of a linear memory, the Transformer possesses a relational one. Conceptually, for each word it processes, the model generates three vectors: a Query (representing the word's request for context: "who am I and what is my role in this sentence?"), a Key (representing the word's own identity or 'label'), and a Value (representing the word's actual meaning).

The magic of self-attention lies in how these three vectors interact. To determine the

relevance of every other word to the word currently in focus, the model calculates the dot product between the *Query* vector of the focus word and the *Key* vector of every other word in the sentence. This operation produces a raw "attention score" for each word pair. These scores are then passed through a softmax function, which normalizes them into a set of weights that sum to one. These weights can be interpreted as percentages, indicating exactly how much "attention" the focus word should pay to each of its neighbors. A high weight means a strong contextual connection. Finally, a new, contextually-enriched vector for the focus word is computed as a weighted sum of all the *Value* vectors in the sentence, using the attention weights just calculated. This process is performed in parallel for every word in the sentence, typically by leveraging highly optimized matrix multiplications.

Crucially, this entire self-attention mechanism is designed as a series of transformations that, through normalization and residual connections, do not alter the fundamental dimensionality of the word vectors. The output of an attention layer is a vector of the same size as the input; what changes profoundly is the information encoded within its numerical components. The initial, static vector representing a word is thus transformed into a dynamic, context-aware representation that has absorbed relevant semantic information from its entire surrounding sentence. This process can be stacked in multiple layers, allowing the model to learn increasingly complex and abstract linguistic relationships.

This technological leap directly addresses the limitations of traditional analytical tools. The standard analytics dashboards provided by social media platforms operate largely on the "bag-of-words" principle: they can count keyword frequencies, track hashtag usage, and measure engagement on a per-post basis. However, they lack the capacity to perform the deep semantic analysis described above. They cannot distinguish between different meanings of the same word, understand sarcasm, or capture the nuanced relationships that form the true meaning of a sentence. Consequently, it would be impossible to obtain the same depth of insight with these classic tools or with less sophisticated models. The Transformer architecture is the indispensable prerequisite for moving from simply counting what is said to truly understanding it. This opens the door to a new generation of language models, which leverage this powerful architecture to build a general, foundational knowledge of human language.

2.2.2 BERT and the Evolution of Pretrained Models

The Transformer architecture provided the necessary hardware, but it was the subsequent development of models like BERT (Bidirectional Encoder Representations from Transformers) that truly unlocked its potential for general-purpose language understanding. BERT's innovation was not in its architecture, it is, in essence, the encoder stack from the original Transformer, but in its revolutionary training methodology Devlin et al., 2019.

BERT introduced the concept of pre-training, a process where the model is trained on a colossal amount of unlabeled text data (billions of words from sources like Wikipedia) with the sole objective of "understanding language" itself, before being applied to any specific task. This is a fundamental departure from earlier models, which had to be trained from scratch for each individual problem. This pre-training is accomplished through two ingenious, unsupervised tasks. The first is the Masked Language Model (MLM) task, where approximately 15% of the words in a sentence are randomly hidden (or "masked"), and the model's goal is to predict these hidden words based on the surrounding, unmasked

context. Because the Transformer's self-attention mechanism looks at the entire sentence at once, the model must learn deep, bidirectional relationships to succeed at this task. It must understand grammar, syntax, and semantics to infer the missing piece. The second task is Next Sentence Prediction (NSP), where the model is given two sentences and must determine if the second sentence is the one that naturally follows the first in the original text, or if it is just a random sentence from the corpus. This forces the model to learn about the logical cohesion and relationship between sentences.

The result of this intensive and computationally expensive pre-training is a versatile, general-purpose language "engine." This pre-trained model is not a final product, but a foundational layer of deep linguistic knowledge. Crucially, this is the same underlying technology that powers the sentiment analysis models used in this thesis. The key advantage for a platform like Viralba is that it does not need to train these massive models from the ground up. Instead, it can take a pre-trained model like BERT (or its multilingual variant, XLM-Roberta) and perform a much faster and cheaper process called fine-tuning. This involves continuing the training on a smaller, task-specific dataset—for instance, a dataset of tweets labeled for sentiment.

This two-stage process (pre-training and fine-tuning) is what makes these models so powerful and superior to less sophisticated approaches. The analytics tools provided by social media platforms can count keywords, but they cannot perform this kind of deep, contextual analysis. They lack the foundational linguistic knowledge that BERT acquires during pre-training. Consequently, they fail at the very tasks, understanding nuance, sarcasm, and complex sentence structures, where these models excel. This capability gap is what necessitates the use of advanced AI models and forms the core of Viralba's value proposition: providing a level of understanding that was previously unattainable.

2.3 Measuring Tone in Social Media Conversation

With the foundational technology of the Transformer and pre-trained models like BERT established, we can now explore its first practical application within the Viralba platform: understanding the affective tone of the conversation. The ability to automatically gauge how a community feels, whether the sentiment towards a brand is positive or negative, or whether a new product launch is met with joy or disappointment, is a cornerstone of modern social intelligence. This capability is realized through two related but distinct analytical tasks: Sentiment Analysis and the more granular Emotional Analysis. The bridge between the general-purpose language "engine" described previously and these specific, value-driven applications is a process known as fine-tuning.

2.3.1 Fine-Tuning for Sentiment and Emotion Analysis

A pre-trained model like BERT or its variants possesses a deep, general-purpose understanding of language, but it does not, out of the box, know how to perform a specific task like classifying sentiment. It has learned the grammar, semantics, and relationships between words, but it has not yet been taught to apply this knowledge to a concrete business problem. The process of adapting this general engine to a specialized skill is known as fine-tuning.

Conceptually, fine-tuning involves taking the entire pre-trained model, with its billions of

learned parameters, and adding a new, small layer of neurons on top of it. This new layer, often called a "classification head," is initially untrained and its parameters are randomly initialized. The entire combined model is then further trained, but this time on a much smaller, task-specific dataset that has been manually labeled by humans. For the task of sentiment analysis, this would be a dataset composed of several thousand tweets, each carefully labeled by an annotator as positive, negative, or neutral.

During this fine-tuning phase, the model's pre-existing knowledge is not erased. Instead, as the model learns to perform the new classification task, its deep linguistic understanding is gently "steered" or "adapted" to focus on the specific patterns, nuances, and keyword indicators that are relevant to sentiment. This two-stage approach, extensive pre-training on unlabeled data followed by brief fine-tuning on labeled data, is incredibly efficient. It leverages the vast knowledge acquired from the billions of sentences seen during pre-training, allowing the model to achieve state-of-the-art performance on a new task with only a fraction of the data and computational cost that would be required to train a model of this size from scratch.

2.3.2 Adapting XLM-T for Viralba's Sentiment Detection

For a platform like Viralba, which must analyze the unique and often chaotic language of social media, choosing the right pre-trained model to fine-tune is a critical strategic decision. As previously discussed, the informal nature of online discourse, with its reliance on slang, emojis, and non-standard grammar, presents a significant challenge. A generic language model, pre-trained on clean and formal text like Wikipedia, would struggle to capture the true sentiment behind this type of language.

It is for this reason that Viralba's analytical pipeline is built upon a specialized, domain-adapted model: XLM-T Barbieri et al., 2022. This model's strength lies in its tailored training process. It begins with a powerful, multilingual pre-trained base, XLM-Roberta, which already possesses a broad understanding of over a hundred languages. Crucially, it then undergoes an additional pre-training phase, a process known as domain adaptation, on a massive corpus of nearly 200 million tweets. By being exposed to this vast quantity of real-world social media text, the model learns the specific nuances, syntax, and vocabulary that characterize these platforms.

By starting from this Twitter-native foundation, the subsequent fine-tuning for sentiment analysis becomes far more effective and reliable. The model does not need to learn the meaning of an emoji from scratch; it has already seen it used in millions of different contexts. When fine-tuned on a labeled sentiment dataset, it can more easily and accurately learn to associate specific linguistic patterns with a *positive*, *negative*, or *neutral* polarity.

The business implication of this enhanced accuracy is the ability for Viralba to offer its clients a robust and trustworthy tool for real-time brand health monitoring. A brand can use this to track sentiment fluctuations with high confidence, measuring the immediate public reaction to a marketing campaign, a new product, or a public relations statement. This capability transforms sentiment analysis from a simple academic exercise into a vital tool for strategic decision-making, allowing companies to understand the impact of their actions as they unfold.

2.3.3 Capturing Emotions Behind User Reactions

While a simple positive-negative polarity provides a valuable high-level overview, it often fails to capture the full spectrum of human expression and the specific drivers behind user sentiment. A more granular and strategically powerful approach, also central to Viralba's offering, is Emotional Analysis. This technique moves beyond the coarse categorization of polarity to classify text into a richer set of discrete emotional categories, such as *joy, anger, sadness, fear*, or *surprise*.

The underlying technology is identical to that of Sentiment Analysis: a domain-adapted, pre-trained model like XLM-T is fine-tuned, but this time on a dataset that has been labeled by human annotators with these more specific emotional tags. While the technical process is similar, the strategic value of the output is exponentially greater.

Knowing that 30% of user mentions are 'negative' is a generic alert; it signals a problem but provides no information about its nature. Knowing that 25% of mentions express 'anger' while 5% express 'sadness' is a diagnosis. This level of granularity allows a company to act with surgical precision. 'Anger' is often a reaction to a functional failure: a service outage, a faulty product, a poor customer service experience and requires an immediate, decisive response from operational or support teams to mitigate the damage. 'Sadness', on the other hand, is frequently a response to a product's failure to meet expectations or the discontinuation of a beloved feature; this feedback is less urgent but critically important for the product development and strategy teams.

Emotional Analysis thus transforms a vague metric into a precise, departmental-level call to action. It provides not just data, but a diagnosis of the community's affective state, enabling Viralba's clients to respond to feedback with the appropriate tone, resources, and strategic focus, turning a potential crisis into an opportunity for improvement and deeper customer engagement.

2.4 Detecting Conversation Topics with BERTopic

While Sentiment and Emotional Analysis provide a crucial understanding of the affective tone of a community, they do not address the fundamental question of *what* is being discussed. To uncover these themes, a platform like Viralba must employ a robust Topic Detection algorithm. The goal is to automatically scan through thousands of documents and identify the main underlying topics of conversation, a task for which Viralba has developed a sophisticated, multi-stage pipeline built around the state-of-the-art BERTopic model.

2.4.1 Data Cleaning and Preparation for Topic Modeling

The first stage of Viralba's pipeline is a rigorous data preparation and filtering process. The quality of a topic model is critically dependent on the quality of the input data; therefore, a series of pre-processing steps are applied to clean and normalize the raw text. This includes the normalization of hashtags (e.g., converting #TopicModeling and #topicmodeling to a single format), and the complete removal of non-semantic elements such as user mentions, emojis, and hyperlinks, which could otherwise introduce noise into the analysis.

Following this initial cleaning, a crucial filtering step is performed. To ensure that the topic model is built upon sufficiently rich textual content, all documents that do not meet a minimum length requirement (specifically, at least ten words post-processing) are discarded. This prevents very short or trivial texts, such as one-word replies, from being included in the analysis, thereby improving the coherence and stability of the resulting topics.

2.4.2 From Sentence Embeddings to Topic Clusters

Once the data has been cleaned and filtered, the core topic modeling process begins. This phase leverages the BERTopic framework and can be broken down into three distinct steps.

First, the pre-processed documents are subjected to vectorization. Using a pre-trained language model based on the Transformer architecture, each document is converted into a numerical vector, or "embedding," that semantically represents its content. As detailed in the previous section, this ensures that texts with similar meanings are located in proximity to one another in a high-dimensional vector space.

The second step is clustering. The collection of document vectors is then processed by the HDBSCAN algorithm, a powerful density-based clustering method. This algorithm groups the vectors based on semantic similarity, effectively aggregating the documents into dynamic clusters, each representing a potential topic. A key feature of HDBSCAN is its ability to identify outliers, in this context, documents that do not appear to belong to any coherent topic. These are temporarily set aside, ensuring that the initial topic clusters are as cohesive as possible.

The third and final step in this core phase is keyword extraction. For each identified cluster, the top ten most representative keywords are calculated. This is achieved using an innovative modification of a classic information retrieval formula known as TF-IDF (Term Frequency-Inverse Document Frequency). The standard TF-IDF is designed to measure the importance of a word within a single document relative to a whole collection of documents by combining two metrics: Term Frequency (TF) and Inverse Document Frequency (IDF). The Term Frequency is a simple measure of how often a word appears in a given document; a high frequency suggests importance. The Inverse Document Frequency, which provides the core insight, measures how common or rare a word is across all documents in the collection. The IDF score is high for rare words and low for very common words, based on the logic that terms that are frequent in a specific document but rare overall are the most descriptive and uniquely representative of that document's content.

BERTopic cleverly adapts this logic to the context of topics instead of documents. Its class-based TF-IDF (c-TF-IDF) first treats all texts within a single cluster (a "class" or topic) as one large, composite document. It then calculates the importance of each word not based on its rarity across documents, but based on its rarity across *topics*. In doing so, it identifies the keywords that are most uniquely characteristic of a given topic cluster. Before this calculation, a set of generic stopwords is removed to ensure the results are semantically meaningful. The output of this stage is an initial, raw version of the topics, each represented by a list of ten keywords.

2.4.3 How Viralba Optimizes and Labels Topics with AI

The procedure described thus far yields a solid first version of the topics present in the data. However, the raw output of this core process often presents two significant limitations that can hinder its direct application in a business context. First, a considerable number of documents may be labeled as outliers by the HDBSCAN algorithm. While this is a feature that ensures the high coherence of the core topics, it can also mean that a substantial portion of the data remains unclassified. Second, the number of generated topics can be particularly high, with several topics often being semantically very similar and thus conceptually redundant.

To overcome these limitations and transform the raw output into refined, strategic insights, Viralba's pipeline introduces a sophisticated optimization and labeling layer. This post-processing stage is where the platform adds its most significant value, moving beyond the standard BERTopic implementation.

The first step in this refinement phase is outlier reduction. The pipeline systematically attempts to reduce the number of unclassified documents by analyzing their semantic similarity to the established topic clusters. For each outlier document, the model calculates its proximity to the existing topics. If a document's embedding is sufficiently close to a specific cluster, it is re-assigned from the "outlier" category to that topic. This is followed by an update of the topic's representative keywords to reflect its new composition.

The second step is topic optimization through aggregation. To address the issue of redundant and overly similar topics, the system calculates a similarity matrix between all generated topic representations. It then iteratively merges pairs of topics whose semantic cosine similarity exceeds a predefined threshold (set at 0.85). When two topics are merged, a new, combined set of keywords is calculated for the resulting aggregated topic. This automated process consolidates fragmented themes into more robust and coherent high-level topics, significantly improving the clarity and usability of the final output.

The final and most innovative step is automated topic labeling via a Large Language Model (LLM). The "raw" output of a topic model is a list of keywords, which, while representative, are not always immediately "human-readable" and require interpretation. To bridge this final gap between data and insight, Viralba's system leverages an LLM. The list of top keywords for each optimized topic is passed to the LLM, which is tasked with interpreting these terms and generating a concise, descriptive title for the topic. This process can even trigger a final optimization loop: if the LLM, with its deep semantic understanding, determines that two different sets of keywords refer to the same underlying concept, it can assign them the same label, effectively flagging them for a final merge. This last step completes the transformation, converting a raw list of statistically relevant words into a fully-fledged, interpretable, and strategically valuable business insight.

2.5 Bringing Sentiment and Topics Together

The methodologies detailed in this chapter, Sentiment Analysis, Emotional Analysis, and contextual Topic Detection, form a powerful analytical triad. Their combination allows a platform like Viralba to move beyond simple data aggregation and to build a rich, multi-dimensional understanding of the content of online conversations. By deploying these algorithms in concert, it becomes possible to answer not only *"how"* a community feels through its tone and emotion, but also *"what"* it is feeling about, by identifying the specific topics driving those affective responses. This synergy provides a robust foundation for data-driven business intelligence, enabling companies to monitor brand health, diagnose customer issues, and identify thematic trends with unprecedented speed and accuracy.

However, a content-only approach, no matter how sophisticated, provides an incomplete picture of a community. It treats the online environment as a collection of disembodied texts, ignoring the relational fabric that connects them. An analysis of content alone leaves critical strategic questions unanswered: Who are the key voices driving these conversations? Are the negative sentiments about a new feature originating from a few highly influential users or from a broad, diffuse base of customers? How do new ideas and opinions propagate through the network? Answering these questions requires a shift in analytical perspective, from the content of the messages to the structure of the relationships between the authors.

This fundamental gap is precisely what Social Network Analysis (SNA) is designed to address. While the tools discussed thus far analyze the text, SNA analyzes the connections. The following chapter will introduce this methodology as the final, strategic extension to Viralba's analytical framework, demonstrating how it can be used to map the architecture of a brand's community and uncover the dynamics of influence that a purely content-based analysis would inevitably miss.

Chapter 3

Testing the Potential of Network Analysis for Brands

3.1 Extending Viralba with Network Analysis

Building upon the applications proposed by Viralba, this chapter will focus on presenting the startup's next strategic objective: the integration of Network Analysis alongside the already developed algorithms presented in Chapter 2. Network Analysis serves as a powerful instrument for gaining a profound understanding of the structure and relationships within a given network. The purpose of this chapter is to simulate an analysis that could potentially be introduced and then performed by Viralba, studying the network of a specific company to demonstrate the profound potential of this type of study.

This objective will be pursued by addressing the following primary research question: *To what extent can network analysis delineate the community structure of a brand's audience, identify its constituent segments (communities), and pinpoint potential or actual micro-influencers who hold significant relevance within the network and are capable of influencing brand perception among a subset of users?*

To test this analytical approach, the sports streaming service DAZN was selected as a case study. DAZN represents an ideal subject as it fulfills two essential criteria for this analysis. Firstly, it maintains a significant and active presence on social media platforms. Secondly, it is frequently at the center of public debate and discussion. This latter point is particularly crucial; conducting this analysis with simpler tools than those Viralba would eventually develop requires a subject rich in debate, controversy, or discussion. Choosing a company with a limited social presence or one that is rarely the subject of public discourse would risk compromising the successful identification of micro-influencers or well-defined topic-based communities.

3.2 Collecting and Structuring the Dataset for Social Graphs

The foundational step for this quantitative analysis was the collection of data. The choice of social media platform was a critical decision, and X (formerly Twitter) was selected

due to its inherent suitability for Social Network Analysis. The nature of X, where news is rapidly disseminated and often becomes the subject of extensive commentary and user-to-user conversation threads, makes it a fertile ground for this research. It is arguably superior in this context to other social networks whose primary function may be the diffusion of visual content, such as images or videos, even if they also possess comment sections. The text-centric and public conversational flow of X provides a richer source of relational data.

To obtain the necessary data, a pre-built web scraping actor, `apidojo/tweet-scraper` was used via the Apify platform. This tool proved particularly effective due to its high degree of customizability, allowing for the precise setting of search parameters such as keywords, dates, language, and other metadata. The scraper delivered its output in either Excel or JSON format, facilitating the subsequent selection of relevant data columns for the analysis.

To initiate the analysis, the search term was set simply to "DAZN" in order to capture any form of interaction with the brand, whether DAZN was directly mentioned, included as a hashtag, or merely appeared in the text of a tweet. Due to the technical limitations of the scraper, the collection process was iterated multiple times to obtain several datasets, each containing between 6,000 and 8,000 tweets. These datasets were subsequently combined to form a single, comprehensive corpus totaling 25,849 tweets. During these iterations, the date ranges were varied to ensure a broad yet focused capture of the summer timeline, a period of heightened interest for the brand.

The combination of these datasets revealed a non-continuous data collection period, with a notable overlap during the month of July. This overlap necessitated a rigorous deduplication process to ensure the integrity of the final analytical dataset.

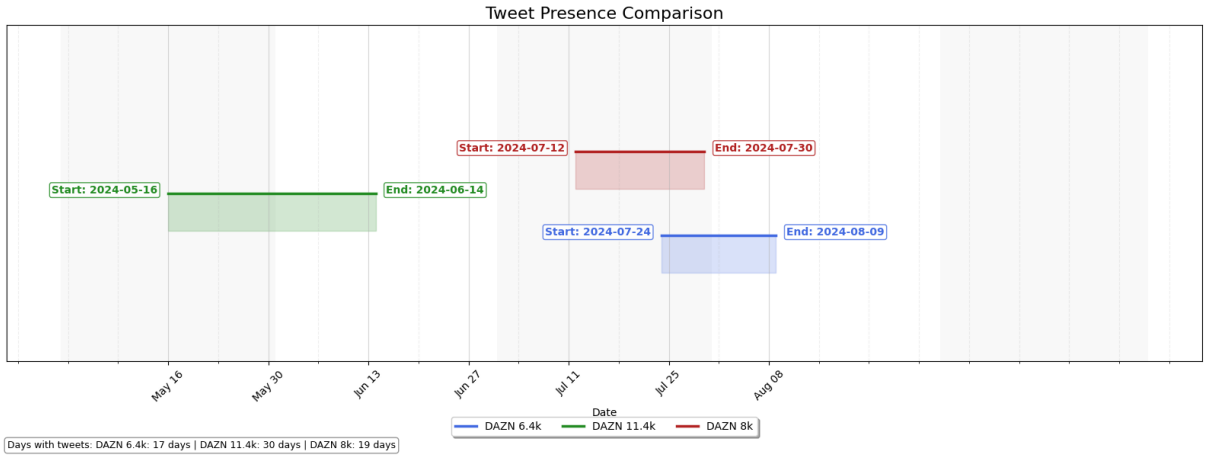

Figure 3.1: Temporal Distribution of the Raw Data Scrapes. The chart illustrates the collection periods of the three initial datasets, highlighting their overlap in July and the resulting composite timeline for the study.

To avoid distorting the analysis with redundant data, duplicates were eliminated by iterating through the dataset and retaining only entries with a unique tweet `id`. This procedure resulted in a final, cleaned dataset of 22,980 unique tweets.

The subsequent step involved formatting this dataset into a structure legible by Gephi,

the software chosen for this analysis. Gephi is an open-source software platform for the analysis and visualization of complex networks and graphs. Widely used in fields such as social science, biology, marketing, and information security, it allows for the exploration of network structures through interactive representations and advanced metrics. Its primary function is to transform large, relational datasets into comprehensible graphs, making it possible to easily identify central nodes, communities, and hidden patterns. Gephi's strengths lie in its intuitive interface, its capacity to handle large-scale networks in real-time, and its rich library of integrated algorithms for topological analysis and community detection. These features, combined with its extensive graphic customization options, make it a powerful tool not only for analysis but also for the visual communication of data. For this study, the data was structured into an adjacency list format for the user-mention network. The columns containing explicit user mentions, labeled `mention/0/text` through `mention/9/text`, were extracted alongside the tweet's author (`author/username`) to form the source and target nodes for the network's edges.

3.3 Analyzing the User Network Around DAZN

Upon loading the adjacency list into Gephi, the first network graph was generated. To render the network's structure in a clear and intuitive manner, the ForceAtlas2 layout was applied. This is one of the most widely used layout algorithms in Gephi, based on a force-directed principle. It simulates a system of physical forces, where an attractive force pulls connected nodes together (as if joined by springs) and a repulsive force pushes all nodes apart (as if they were electrically charged). The result is a spatial arrangement where more connected nodes tend to cluster in the center, while less connected ones are distributed towards the periphery. ForceAtlas2 is particularly effective for visualizing communities or cohesive groups within a network, as it naturally separates dense areas from sparser ones. It is ideal for medium-to-large-scale networks, particularly for tasks such as social media analysis, where identifying clusters and central nodes is crucial.

The visual representation of the network was further enhanced by coloring the nodes according to the communities detected and by sizing them based on their influence scores.

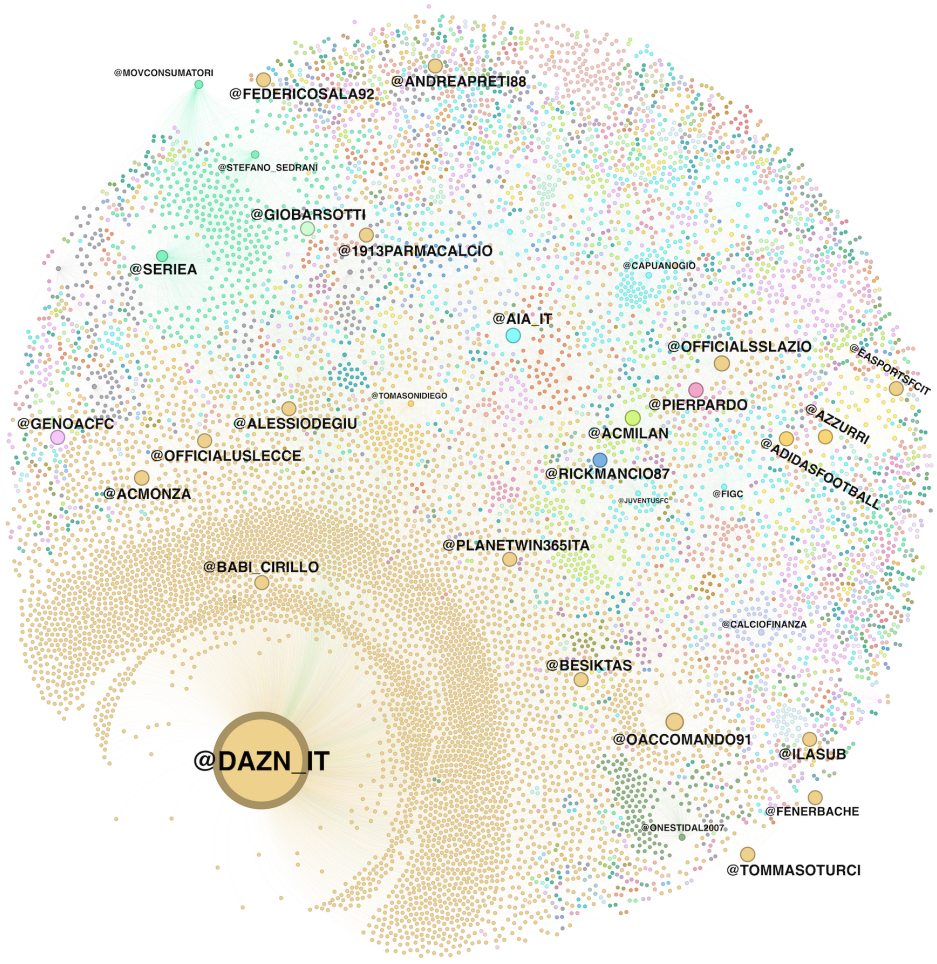

Figure 3.2: Visualization of the General DAZN User-Mention Network. The network is spatially organized by the ForceAtlas2 algorithm, with nodes sized by PageRank and colored by community (Modularity class). The structure reveals a dense core and distinct peripheral communities.

3.3.1 Algorithms for Calculating Influence and Detecting Communities

The visual architecture of the network shown in Figure 3.2 provides powerful at-a-glance insights into the community's structure. However, to systematically decode the complex patterns of influence and segmentation, a deeper understanding of the quantitative methods that generated this visualization is required. The following sections therefore provide a conceptual overview of the three foundational algorithms employed in this analysis to measure influence and detect communities.

Modularity for Community Detection

The primary algorithm used to identify communities within the network was Modularity. This is not simply a procedure, but a quality score that evaluates how well a network is partitioned into different clusters. Its core principle is to compare the given structure to a "null model"—a hypothetical random network with the same number of nodes and the same degree distribution Newman, 2006. A community structure is considered statistically significant if the number of edges falling *within* the identified communities is substantially higher than the number one would expect to find by random chance. A high modularity score, therefore, signifies a strong, non-random community structure where internal connections are far denser than external ones. The algorithm implemented in Gephi iteratively optimizes the network's partitions to find the division that maximizes this modularity score, which is then used to assign a unique color to each identified group.

PageRank for Measuring Authority

To measure the influence of each user, the primary metric was PageRank. Originally developed to rank web pages, its logic models a flow of "prestige" or "authority" through a network Page et al., 1999. The core idea is that a node's importance is determined not only by the number of incoming links (mentions) it receives but, crucially, by the importance of the nodes that are the source of those links. Conceptually, it models a "random surfer" navigating the network: the PageRank score of a user represents the long-term probability of this surfer landing on that user's node. This makes PageRank a robust metric for identifying truly authoritative figures, as a mention from a highly influential account confers more "rank" than a mention from a peripheral one.

Eigenvector Centrality for Measuring Local Influence

Operating on a similar principle of recursive influence, Eigenvector Centrality posits that a node's centrality is a function of the centrality of its neighbors Bonacich, 1972. A node will have a high eigenvector score if it is connected to many other nodes that themselves have high scores. While PageRank measures a more global form of authority, Eigenvector Centrality is highly effective at identifying influential nodes within dense, cohesive clusters—the most important members of the most important "clubs" in the network. In this analysis, it served as a complementary metric to PageRank, helping to validate findings and provide additional insight into the local prominence of community leaders.

3.3.2 Institutional and Official Accounts at the Core of the Network

The initial analysis of the network's key actors, based on PageRank scores, reveals a clear and stratified hierarchy of influence. As expected, the official brand account, @DAZN_IT, functions as the gravitational center of the entire ecosystem. It holds the highest scores across all major centrality metrics, confirming its role as the primary target of user mentions and the central hub of the conversation.

However, a deeper look at the top-ranking nodes reveals that the sphere of influence extends beyond the brand itself, being primarily composed of two distinct, yet interconnected, groups. The first can be defined as the "Institutional Sphere," comprising the official accounts of major football clubs (e.g., @ACMILAN, @OFFICIALSSLAZIO), national and international football organizations (@AZZURRI, @SERIEA), and major commercial partners (@ADIDASFOOTBALL). The high centrality of these nodes is structurally logical, as they represent the core content and context within which DAZN operates. They are frequently co-mentioned with DAZN in discussions about specific matches, broadcasting rights, and league-wide news.

The second dominant group is the "Official Sphere of Influence," which consists of individuals professionally affiliated with DAZN. This cohort includes prominent commentators, journalists, and on-air personalities who act as the public faces of the brand, such as @OACCOMANDO91, @PIERPARDO, and @TOMMASOTURCI. Their high PageRank scores are a testament to DAZN's effective communication strategy, which leverages its talent as key amplifiers and conversational hubs. They serve a dual function: broadcasting official information and acting as accessible targets for the community's questions and comments.

The combined dominance of these two spheres demonstrates that the most visible layer of the DAZN network is characterized by a top-down flow of information. The most central and authoritative voices are those directly or indirectly affiliated with the brand and its core business. While this finding confirms a predictable and well-organized communication structure, the primary objective of this research was to penetrate this surface layer to identify more authentic, grassroots sources of influence.

3.3.3 Finding Independent Voices Through Strategic Filtering

To move beyond the official and institutional voices that dominate the network's surface, a multi-stage filtering methodology was designed and implemented using a Python script with the `pandas` and `networkx` libraries. The goal was to systematically isolate a cohort of users who are both quantitatively influential and qualitatively independent from the DAZN brand. This process transforms the concept of a "micro-influencer" from a vague notion into a methodologically defined and replicable classification.

The process began by establishing a baseline of influence. All nodes in the network were ranked by their PageRank score, and only those ranking in the 98th percentile or higher were retained for further analysis. This initial step significantly reduced the number of nodes, focusing the investigation on the top 2% most authoritative actors in the ecosystem.

Secondly, a qualitative filtering step was applied by creating a comprehensive "exclusion list." This list was manually populated with the usernames of all previously identified in-

stitutional accounts (football clubs, leagues, brands) and official brand affiliates (DAZN's own channels, known commentators, and media personalities). This list was then used to remove these accounts from the pool of influential nodes, a crucial step designed to separate brand-driven influence from authentic community-level influence.

Finally, to mitigate the PageRank algorithm's known anomaly whereby a single mention from a highly authoritative node can disproportionately inflate a user's score, as we will see later, a minimum interaction threshold was established. Only users who had received a minimum of 10 incoming mentions (In-Degree ≥ 10) were included in the final list. This ensures that the identified micro-influencers are not just passive recipients of a high-profile mention but are active and sustained participants in the community's conversations.

This rigorous filtering process yielded a final, curated list of independent micro-influencers. The top fifteen of these actors are presented in the table below, ranked by their PageRank score.

Table 3.1: Top 15 Identified Independent Micro-Influencers in the General DAZN Network, with Associated Centrality Metrics.

Rank	User	PageRank	In-Degree	Betweenness	Eigenvector Centrality
1	@AUGUSTOCIARDI75	0.00388	51	0.00013	0.00785
2	@ONESTIDAL2007	0.00373	205	0.00328	0.00868
3	@MOVCONSUMATORI	0.00355	350	0.00000	0.02056
4	@STEFANO_SEDRANI	0.00289	346	0.00004	0.01826
5	@TOMASONIDIEGO	0.00285	183	0.00305	0.00141
6	@NONEVOLUTO	0.00166	107	0.00123	0.00009
7	@SPORTFACE2016	0.00135	87	0.00000	0.00031
8	@IL_BILLA	0.00132	63	0.00288	0.01520
9	@FABRAVEZZANI	0.00131	29	0.00000	0.00099
10	@PIERRE8244309	0.00129	54	0.00644	0.00092
11	@PAP1PAP	0.00127	86	0.00125	0.00029
12	@FRALITTERA	0.00121	56	0.00242	0.00279
13	@ADLMAIALE	0.00117	21	0.00384	0.00021
14	@DANIELEBIBO	0.00112	72	0.00003	0.00927
15	@LAPONERO	0.00107	66	0.00001	0.00014

3.3.4 Understanding the Roles of Key Users in the Network

The resulting list of users is markedly different from the initial, unfiltered ranking. It reveals a diverse set of actors who hold significant sway within the community despite having no direct connection to DAZN. An analysis of their roles, interpreted through their unique combination of centrality scores, allows for their categorization into distinct archetypes of influence. A close examination of the centrality metrics associated with the users in Table 3.1 reveals that "influence" is not a monolithic concept. Instead, different actors perform distinct structural roles within the network. Based on their statistical profiles, we can identify at least four key archetypes of independent influencers.

The first and most prominent archetype is the Community Leader. These are users who

serve as central hubs for specific fan communities. Representative examples from the analysis include @ONESTIDAL2007 (a Juventus-focused parody account) and @STEFANO_SEDRANI (a passionate Udinese supporter). Their defining characteristic is an exceptionally high In-Degree score (205 and 346, respectively), indicating that they are frequent targets of mentions and central points of reference for their follower base. Their influence is deep and authoritative within their niche. While their PageRank is high, their Betweenness Centrality is relatively modest, suggesting they are powerful centers of their own communities but may not necessarily serve as bridges to others. They are the "tribal chieftains" of the DAZN ecosystem.

The second archetype is the Information Broker. These users are structurally crucial for the flow of conversation across the network. Their defining metric is a high Betweenness Centrality score, which signifies that they frequently lie on the shortest communication paths between otherwise disconnected users or communities. Accounts such as @TOMASONIDIEGO, @PIERRE8244309, and @IL_BILLA exemplify this role. While their In-Degree may be lower than that of Community Leaders, their function as connectors is vital. They prevent the network from fracturing into isolated echo chambers and are essential for the diffusion of information and opinions across different fan groups. They are the "network's glue."

A third, particularly interesting archetype is the Institutional Watchdog. The most striking example of this is @MOVCONSUMATORI (an Italian consumer rights association). This account possesses the highest In-Degree (350) among all independent influencers, yet its Betweenness Centrality is zero. This unique statistical signature indicates that it does not actively participate in brokering conversations. Instead, it is massively "invoked" or "tagged" by users, likely in discussions related to service issues, pricing, and consumer rights. It functions as a "court of appeal" or an external authority to which users direct their grievances. Its presence and high rank demonstrate that the conversation around DAZN frequently transcends sport and enters the domain of consumer advocacy.

Finally, the fourth archetype identified is that of the Independent Media and Pundit. This category includes accounts like @AUGUSTOCIARDI75 and @SPORTFACE2016, which belong to journalists or smaller media outlets not officially affiliated with DAZN. These actors maintain a solid balance across all centrality metrics, indicating that they are not only popular (In-Degree) and authoritative (PageRank) but also serve, to some extent, as information brokers. They represent a "third voice" in the ecosystem, providing news, commentary, and opinions that are perceived as being independent from both the official brand narrative and the pure fan-driven discourse. They are the alternative sources of information to which the community turns for external validation and analysis.

The identification of these distinct archetypes fulfills one of the primary objectives of this study. It demonstrates that by applying a systematic, multi-stage filtering methodology, it is possible to move beyond a simplistic ranking of "top influencers" and achieve a nuanced, strategic understanding of the different functional roles that key actors play within a complex brand community.

3.4 Exploring Topics Through Hashtags

Having analyzed the structure of user interactions, the research next sought to understand the thematic landscape of the DAZN conversation. A simplified yet effective form of topic

detection was performed by analyzing the co-occurrence of hashtags. This approach is particularly well-suited to the X platform, where the character limit encourages the precise use of hashtags to categorize and add context to a tweet.

To conduct this analysis, the same initial adjacency list was used, but instead of user mentions, the hashtags employed by each user were extracted. For simplification, an edge list was generated that links a user to each individual hashtag they used. This data was then loaded into Gephi to generate a hashtag co-occurrence network. As previously described, this process involves projecting a bipartite user-to-hashtag graph into a single-mode network where nodes are hashtags and edges represent their co-use by common users. The layout was again managed by ForceAtlas2, nodes were colored by their modularity class to identify thematic communities, and their size was determined by their In-Degree, representing their overall usage frequency. The final visualization was filtered to show only the top 40 most frequently used hashtags.

Figure 3.3: Hashtag Co-occurrence Network of the Top 40 Most Used Hashtags. Nodes represent hashtags, sized by usage frequency (In-Degree) and colored by thematic community (Modularity class). The network reveals distinct clusters of related topics.

The resulting hashtag network, shown in Figure 3.3, provides a clear map of the primary themes populating the DAZN discourse. The modularity algorithm successfully parti-

tioned the hashtags into distinct and interpretable thematic communities. The largest and most central cluster, often containing generalist hashtags like #DAZN, #SERIEA, and #CALCIO (Football), represents the core, sports-centric conversation. This community connects various sub-topics, including discussions related to specific football clubs such as #JUVENTUS, #INTER, and #NAPOLI, as well as conversations about the transfer market, indicated by hashtags like #DAZNMERCATO. This cluster signifies the "business as usual" discourse, where users discuss the sporting events and content broadcast on the platform.

However, the most significant finding from this analysis is the emergence of a distinct and tightly-knit community of negative sentiment. As can be observed in the lower-right quadrant of the visualization, the modularity algorithm correctly isolated three specific hashtags: #DISDETTADAZN (Cancel DAZN), #DAZNOUT, and #DISDETTAAFFLITTIVA (Grief-stricken Cancellation). These three nodes form their own unique community, colored in aqua. This is a powerful, data-driven confirmation that user complaints are not merely isolated grievances but constitute a coherent and self-contained thematic cluster. The hashtags within this community are used together to signal a specific type of protest, focused on service cancellation due to dissatisfaction with pricing or quality. The ability of the network analysis to automatically identify and separate this "protest cluster" from the general sports chatter is a prime example of its diagnostic power.

The analysis also reveals other minor thematic groups, such as those related to major international events like #EURO2024 and #PARIS2024, demonstrating that the conversation around DAZN extends to all major sporting events for which it holds broadcasting rights. The structure of the hashtag network thus provides a clear, high-level summary of "what" is being discussed, complementing the previous analysis of "who" is leading the conversation.

3.5 Focusing on the Subscription Debate

To further investigate the dynamics of the community during periods of controversy, a separate, more focused analysis was conducted. This involved the creation of the Subscription-Themed User-Mention Network, built from the dataset of 3,089 tweets containing keywords related to pricing and subscriptions. This approach was designed to isolate the conversation about a specific business issue, allowing for a direct examination of how the network's structure and its hierarchy of influence change when the topic shifts from general sports discussion to a sensitive, service-related matter.

The process mirrored the one used for the general network. The filtered dataset was loaded into Gephi, and the same algorithms were applied: ForceAtlas2 for the layout, and Modularity and PageRank for partition coloring and node sizing, respectively.

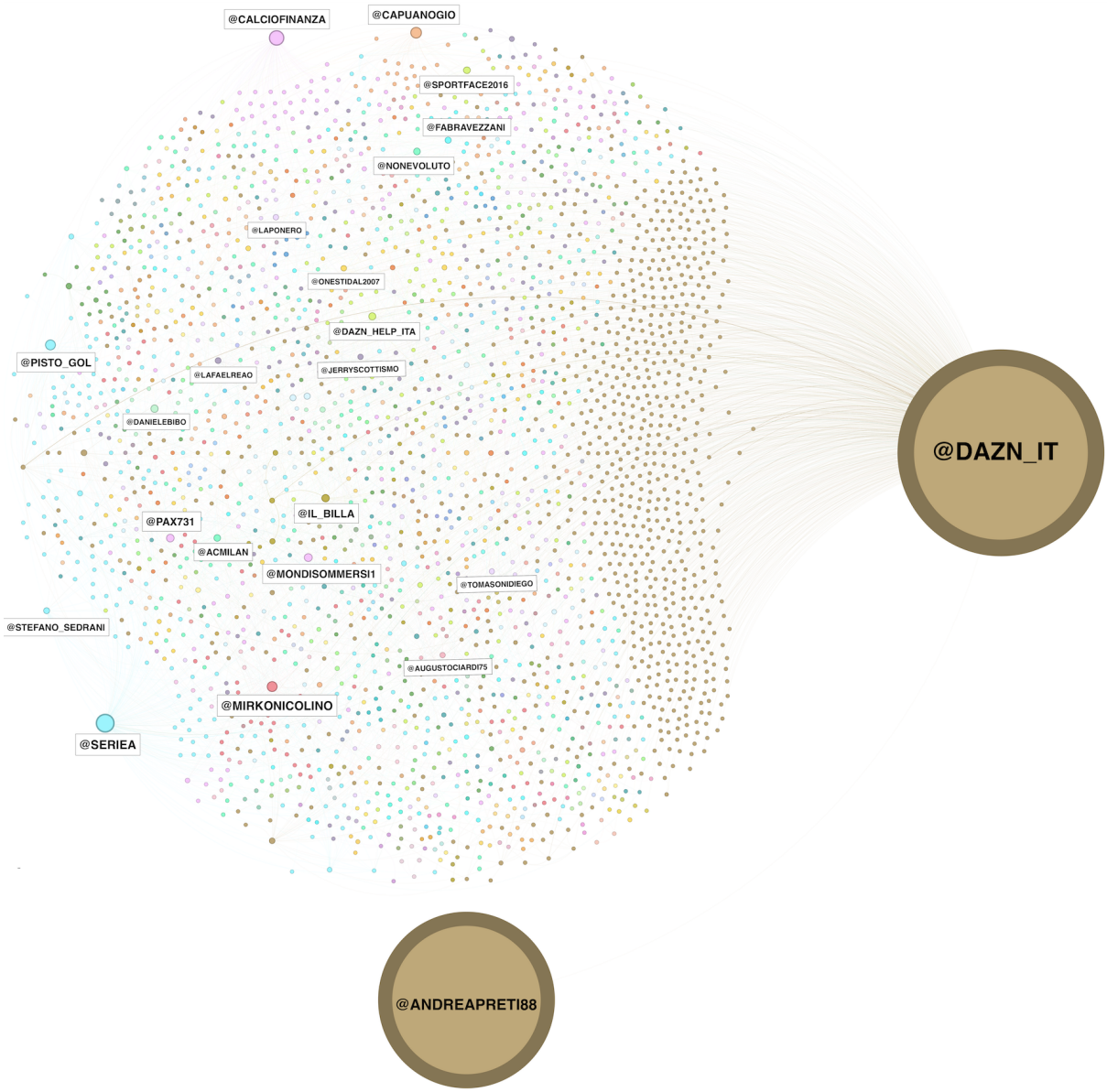

Figure 3.4: Visualization of the Subscription-Themed User-Mention Network. The network illustrates the interaction patterns specific to the conversation around DAZN's pricing and subscription plans. Nodes are sized by PageRank and colored by community.

A striking feature of this network is the high PageRank score of the user @ANDREAPRETI88. A deeper, qualitative investigation revealed this to be an anomaly caused by a specific interaction: the official @DAZN_IT account replied directly to a query from this user, thereby "donating" a significant amount of prestige and inflating their PageRank score. This case serves as an important methodological reminder of how PageRank functions and underscores the necessity of complementing it with other metrics, such as In-Degree, to avoid misinterpreting such anomalies. For this reason, in the subsequent quantitative analysis of this network, the filtering process included a minimum In-Degree threshold to prioritize nodes with sustained community interaction over those with isolated high-profile mentions.

3.5.1 Structural Shifts and the Rise of Specialist Accounts

The true value of this focused analysis emerges when its properties are compared directly with those of the General Network. This comparison reveals a fundamental shift in both the structure of the conversation and, more importantly, in the very definition of what constitutes an influential voice.

First, at a macro-structural level, the Subscription-Themed Network is significantly more cohesive. A quantitative analysis reveals that its network density is 0.00049, a value nearly three times higher than the General Network's density of 0.00017. This empirical evidence supports the hypothesis that a controversial and unifying topic, such as a price increase, causes the engaged community to become more tightly-knit and interactive. The discourse is not only more thematically focused but also structurally more intense.

However, the most profound insight comes from comparing the hierarchies of influence. It is crucial to clarify that the filtering methodology was intentionally and strategically adapted for each network to reflect the changing context of the conversation. For the General Network, a comprehensive exclusion list was applied to filter out all institutional actors, media outlets, and individuals professionally affiliated with DAZN. The aim of this rigorous approach was to isolate the purest form of grassroots, independent community voices.

For the Subscription-Themed Network, this filtering strategy was deliberately modified. While institutional accounts like football clubs were still excluded, the list was adapted to include journalists, media outlets, and DAZN's own on-air talent. This methodological choice was made to test two specific hypotheses: first, to observe whether media personalities would rise to prominence as expert commentators in a business-related debate; and second, to verify whether DAZN-affiliated commentators would actively participate in or remain silent during a sensitive and critical discussion about their employer. This adapted approach allows for a richer analysis of how the entire ecosystem of influence, not just the grassroots layer, reconfigures itself around a specific controversy.

The resulting comparison reveals a dramatic reconfiguration of the influential landscape.

Table 3.2: Comparison of Top-Ranking Actors in the General Network vs. the Subscription-Themed Network.

Rank	Top Actors (General Network)	Top Actors (Subscription Network)
1	@CAPUANOGIO	@CAPUANOGIO
2	@AUGUSTOCIARDI75	@PISTO_GOL
3	@ONESTIDAL2007	@MIRKONICOLINO
4	@MOVCONSUMATORI	@DANIELEBIBO
5	@STEFANO_SEDRANI	@STEFANO_SEDRANI
6	@TOMASONIDIEGO	@SPORTFACE2016
7	@NONEVOLUTO	@NONEVOLUTO
8	@SPORTFACE2016	@FABRAVEZZANI
9	@IL_BILLA	@IL_BILLA
10	@FABRAVEZZANI	@LAFAELREAO
11	@PIERRE8244309	@PAP1PAP
12	@MIRKONICOLINO	@PIERRE8244309
13	@PAP1PAP	@SKYSPORT
14	@FRALITTERA	@ONESTIDAL2007
15	@ADLMAIALE	@LAPONERO

As illustrated in Table 3.2, the nature of the top influencers changes significantly. While the General Network's ranking highlights a diverse mix of grassroots community leaders (e.g., @ONESTIDAL2007, @MOVCONSUMATORI) and brokers, the top of the Subscription Network is almost entirely dominated by journalists and media personalities. Figures like @CAPUANOGIO, @PISTO_GOL, and @MIRKONICOLINO rise to the absolute top of the hierarchy.

This reveals a critical dynamic: during general, fan-centric discourse, influence is distributed among a variety of community archetypes. However, when the conversation shifts to a complex and high-stakes topic like pricing, the community gravitates towards established voices of authority and expertise, the journalists. They become the primary sense-makers and trusted sources, and their influence eclipses that of the fan-community leaders.

Interestingly, several "multi-purpose" influencers, such as @NONEVOLUTO and the community leaders @ONESTIDAL2007 and @STEFANO_SEDRANI, demonstrate their resilience by remaining influential in both contexts, albeit with a lower relative rank in the subscription debate. This confirms their status as robust community hubs, but also underscores the rise of the specialist journalists as the dominant voices during a crisis or controversy. This analysis therefore does not just show a "power shift," but a fundamental change in the type of influence that the community values most, depending entirely on the subject of the conversation.

3.6 Future Applications of Network Analysis in Viralba's Strategy

The network analysis presented in this chapter, while powerful in its own right, represents only one dimension of a truly comprehensive community intelligence framework. The insights derived from understanding the "who" (the actors) and the "what" (the topics) can be exponentially amplified by integrating them with advanced techniques that analyze the "how" and "why" of user sentiment and expression. The future development of Viralba's analytical suite is envisioned to combine network analysis with multi-dimensional Sentiment, Emotional, and advanced Topic Detection analysis, creating a holistic and deeply insightful platform.

One of the most promising avenues is the fusion of network structure with multi-dimensional sentiment and emotional analysis. Instead of a simple positive/negative/neutral classification, this approach can identify a wider spectrum of emotions such as anger, joy, surprise, or disgust within the tweets. By overlaying this emotional data onto the network graph, it would be possible to visualize not just communities, but "emotional hotspots." For instance, one could identify if a specific community cluster is predominantly characterized by anger (e.g., in a price increase discussion) or joy (e.g., after a major sporting victory). This would allow a brand to move beyond merely tracking sentiment scores and begin to understand the specific emotional drivers of different community segments, enabling far more empathetic and effective communication strategies.

Furthermore, the simplified hashtag-based topic detection used in this study can be significantly enhanced by employing more sophisticated models like BERT-based Topic Modeling. Unlike traditional keyword-based methods, these advanced models leverage the contextual understanding of large language models to identify nuanced and emergent topics from the full text of the tweets, even if they don't share common hashtags.

The synergy of these technologies would yield a paradigm shift in business insights. For example, by combining all three, Viralba could answer highly complex and strategic questions:

- Which specific sub-topics, as identified by BERTopic, are driving the most anger within the "Independent Influencer" community?
- Is the "Information Broker" archetype more likely to spread neutral, news-based information or highly emotional content?
- How does the emotional tone of a conversation change as it moves from a central, official node to the periphery of the network?

This integrated approach would transform a static map of the community into a dynamic, multi-layered dashboard of its real-time health, sentiment, and thematic evolution. It would provide businesses not only with a "who's who" of their online ecosystem but a profound understanding of what their community truly thinks, feels, and cares about, moment by moment. This represents the next frontier in data-driven brand management and is the ultimate strategic goal for Viralba's analytical platform.

Conclusion

This thesis set out to explore how an integrated analytical framework could decode the complex influence of online community dynamics to better inform brand and creator strategy. Through a multi-layered investigation, this research has demonstrated the profound limitations of traditional, surface-level metrics and has detailed a more sophisticated, dual-pronged approach that combines the power of Artificial Intelligence for content analysis with Social Network Analysis for structural understanding.

The research presented in the preceding chapters yields several key findings. From Chapter 2, we established that modern AI, built upon the Transformer architecture and models like BERT, provides the necessary tools to move beyond simple keyword counting. It allows for a deep, contextual analysis of *what* a community is discussing through advanced topic detection, and *how* it feels through nuanced sentiment and emotional analysis. From Chapter 3, the application of Social Network Analysis to a real-world case study demonstrated its unique ability to uncover the relational architecture of a community. The analysis successfully identified not only who the most influential actors were, but also categorized them into distinct functional archetypes—such as Community Leaders, Information Brokers, and Institutional Watchdogs. Furthermore, the comparative analysis revealed the fluid, context-dependent nature of influence, showing how the hierarchy of authority shifts dramatically when the topic of conversation changes from general discourse to a specific, high-stakes controversy.

In response to the central research question, this thesis concludes that the strategic integration of AI-based content analysis and Social Network Analysis does indeed provide a powerful framework for generating actionable intelligence. The case study of DAZN empirically demonstrated that this combined approach can successfully decode community dynamics in ways that traditional metrics cannot. It moves beyond "vanity metrics" to identify non-obvious insights, such as the rise of specialist authorities like `@CALCIOFINANZA` during a business-related debate, or the significant structural role of fan-led accounts like `@ONESTIDAL2007`. These are not merely interesting observations; they are concrete, strategic insights that a brand or creator could use to guide their communication, collaboration, and crisis management strategies. The framework, therefore, provides a clear methodological pathway from raw data to deep, actionable understanding.

The implications of this research are significant for both academic study and professional practice. For a platform like Viralba, this thesis serves as a robust proof-of-concept, validating its core technological and strategic vision. For brands and creators, it offers a new playbook for community analysis, urging a shift from a focus on audience size to an understanding of audience structure and thematic sensitivity. For brands, this means a more precise way to identify credible ambassadors and to receive early warnings

on potential crises. For creators, it offers a data-driven method to understand their community more deeply, allowing them to create more resonant content and to identify strategic growth opportunities.

However, this study is not without its limitations. The analysis was conducted on a single case study within a specific cultural and linguistic context (Italian-language discourse around a sports brand). Furthermore, the data was collected over a defined, cross-sectional period and does not capture longitudinal evolutions in the network structure over a longer timeframe. Future research could build upon this framework by applying it to different industries, in different cultural contexts, and through longitudinal studies to track how community structures and influence dynamics evolve over time. Such work would continue to refine our understanding of the powerful, unseen forces that shape our modern digital world. In conclusion, in a world where attention is the scarcest resource, the ability to decode community dynamics is no longer just a competitive advantage, but an essential condition for relevance and success. The tools and approaches outlined in this thesis represent a step towards a future where strategies are not imposed on the market, but are born from a deep and structured listening of the communities themselves.

References

Appel, G., Grewal, L., Hadi, R., & Stephen, A. T. (2020). The future of social media in marketing. *Journal of the Academy of Marketing Science*, *48*, 79–95. https://doi.org/10.1007/s11747-019-00695-1

Ascani, A., & Ancillai, C. (2025). Social media marketing and performance measurement: Does it take two to tango? [(Presumed future publication for thesis context)]. *Journal of Marketing Theory and Practice*.

Barbieri, F., Camacho-Collados, J., Espinosa-Anke, L., & Neves, L. (2022). XLM-T: A multilingual language model for tweet-specific tasks. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 8370–8385.

Barquero Cabrero, M., O'Leary, S., & Schaedler, L. (2023). User-generated content is the most engaging. a study of instagram a/b testing. *Humanities and Social Sciences Communications*, *10*(1), 1–10. https://doi.org/10.1057/s41599-023-01910-6

Bonacich, P. (1972). Factoring and weighting approaches to status scores and clique identification. *Journal of Mathematical Sociology*, *2*(1), 113–120. https://doi.org/10.1080/0022250X.1972.9989806

Deloitte Digital. (2022). Marketing measurement & data – connecting marketing kpis to business objectives [Accessed: 2025-09-11].

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186.

Garhwal, D. S., & Dhanawade, P. D. (2023). Social listening – a review and its use for customer engagement & retention. *Journal of Business and Management*, *25*(3), 44–48.

Hoffman, D. L., & Fodor, M. (2010). Can you measure the roi of your social media marketing? *MIT Sloan Management Review*, *52*(1), 41–49.

HubSpot. (2025). Stop measuring these vanity metrics in your marketing campaign [Accessed: 2025-09-10].

Influencer Marketing Hub. (2022). The state of influencer marketing 2022: Benchmark report [Accessed: 2025-09-06].

JETIR. (2024). Artificial intelligence integration in social media marketing. *Journal of Emerging Technologies and Innovative Research*, *11*(2).

Kemp, S. (2025, February). Digital 2025: The essential guide to the global state of digital [Published by We Are Social. Accessed: 2025-09-15].

Kietzmann, J. H., Hermkens, K., McCarthy, I. P., & Silvestre, B. S. (2011). Social media? get serious! understanding the functional building blocks of social media. *Business Horizons*, *54*(3), 241–251. https://doi.org/10.1016/j.bushor.2011.01.005

Kočišová, K., & Štarchoň, P. (2023). The role of marketing metrics in social media performance evaluation. *Marketing and Management of Innovations*, *14*(1), 1–13. https://doi.org/10.21272/mmi.2023.1-01

Landingi. (2025). Social media marketing vs. influencer marketing: What suits your business best? [Accessed: 2025-09-04].

Liu, W., & Zheng, D. (2024). How influencer-related factors and brand credibility affect purchase intention on social media? the mediating role of brand trust. *Humanities and Social Sciences Communications*, *11*(1), 1–11. https://doi.org/10.1057/s41599-023-01930-2

Matter Communications. (2023). 2023 influencer impact report [Accessed: 2025-09-05].

Newman, M. E. J. (2006). Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, *103*(23), 8577–8582.

Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). *The pagerank citation ranking: Bringing order to the web* (tech. rep. No. 1999-66). Stanford InfoLab.

Rogers, S. (2018). Otherwise engaged: Social media from vanity metrics to critical analytics. *Big Data & Society*, *5*(1). https://doi.org/10.1177/2053951718764089

Schaffer, N. (2025). Content creator vs influencer: What is the difference? [Accessed: 2025-09-02].

Sprout Social. (2024). The social media metrics to track in 2025 (and why) [Accessed: 2025-09-12].

Stieglitz, S., Mirbabaie, M., Ross, B., & Neuberger, C. (2018). Social media analytics – challenges in topic discovery, data collection, and data preparation. *International Journal of Information Management*, *39*, 156–168. https://doi.org/10.1016/j.ijinfomgt.2017.12.002

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, 5998–6008.

Zeng, D., Chen, H., Lusch, R., & Li, S.-H. (2010). Social media analytics and intelligence. *IEEE Intelligent Systems*, *25*(6), 13–16. https://doi.org/10.1109/MIS.2010.151