International Research Journal of Multidisciplinary Scope (IRJMS), 2026; 7(1):241-263 IRJMS

Original Article | ISSN (O): 2582-631X DOI: 10.47857/irjms.2026.v07i01.06486

From Data Mining to Predictive Analytics: Progress in Understanding and Forecasting Social Media Virality

Riovan Styx Roring[1]*, Chih How Bong[2], Narayanan Kulathuramaiyer[2]

[1]Faculty of Science and Technology Information, Kalimantan Institute of Technology, Indonesia, [2]Faculty of Computer Science and Information Technology, University of Malaysia Sarawak, Malaysia. *Corresponding Author's Email: riovan.roring@lecturer.itk.ac.id

Abstract

This paper explores the evolution of data mining techniques into more sophisticated forms of predictive analytics, with a particular focus on their application in understanding and forecasting the virality of social media content within the framework of Society 5.0. The study begins by addressing the role of large-scale data analysis in identifying significant behavioral patterns, correlations, and insights that can serve as early indicators of social media performance. As online interactions have become increasingly dynamic and influenced by numerous social and psychological variables, the research approach has gradually shifted from simple descriptive analytics, which primarily explain past behaviors, toward advanced predictive modeling aimed at forecasting future engagement trends. The methodology integrates a systematic review of social parameters that affect content virality, including factors such as emotional appeal, timing, and user network structures. These social indicators are then examined using data analytics tools capable of enhancing the predictive accuracy of a proposed virality model. By embedding predictive mechanisms into the analytical process, the study enables real-time decision-making in the design and dissemination of digital content. The findings have significant practical implications, particularly for digital marketers, social media strategists, and content creators seeking to optimize their reach and engagement. Ultimately, this research not only presents conceptual framework and methodological progression but also highlights the broader impact of predictive analytics in advancing strategies for digital marketing and enriching the academic discourse on social media research.

Keywords: Content Engagement Analysis, Predictive Analytics, Society 5.0, Social Media Trends, Viral Content Forecasting.

Introduction

The advent of Society 5.0 aims to construct a human-centric society that simultaneously addresses social challenges and fosters economic development, ensuring that individuals can lead active, comfortable, and high-quality lives. This transformative vision is underpinned by a commitment to meet the diverse needs of people across regions, age groups, genders, and languages through the provision of goods and services enabled by advanced technologies such as big data, artificial intelligence, and machine learning (1). Society 5.0 represents a paradigm shift that transcends the limitations of the information society by positioning technology as a means to enhance human welfare rather than an end in itself. Within this context, predictive analytics emerges as a crucial tool for optimizing decision-making and anticipating future needs and trends. Figure 1 shows Society 5.0 Sustainable Development Goals. Society 5.0 is also closely aligned with the United Nations' 17 Sustainable Development Goals (SDGs), emphasizing economic growth, social inclusion, and environmental sustainability. These global objectives encourage the development of smart technologies to bridge existing gaps in intelligence, capability, income, gender, and social status, thereby enabling equal participation in the digital economy (2). As industries and governments adopt data-driven systems to solve complex challenges, predictive analytics plays a pivotal role in translating vast and diverse datasets into actionable insights. The ability to detect behavioral patterns, forecast outcomes, and simulate decision pathways allows organizations to proactively address societal issues while maintaining efficiency and inclusiveness. In this transformative era, the influence of social media platforms has become increasingly significant, serving not only as channels for interpersonal communication but also as engines for information dissemination, marketing, and public opinion formation.

(Received 25th June 2025; Accepted 15th October 2025; Published 10th January 2026)

The explosive growth of user-generated content and digital interactions has generated an abundance of behavioral data that can be mined to reveal the mechanisms of virality—why certain content spreads rapidly and influences mass behavior while other content remains unnoticed. The complexity of these phenomena requires approaches that go beyond traditional descriptive analytics. Predictive analytics, when integrated with data mining techniques, provides the analytical power to forecast trends and user engagement in real time. This evolution from descriptive to predictive analysis marks a major advancement in understanding the dynamics of online influence.

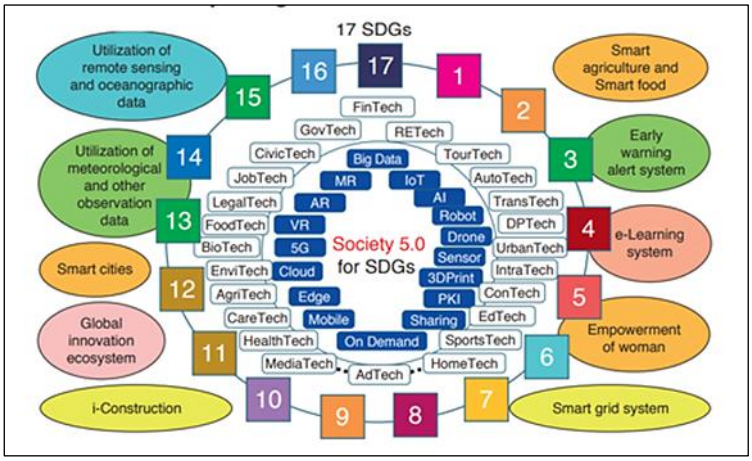

Figure 1: Society 5.0 Sustainable Development Goals (1)

Previous studies have extensively examined the determinants of social media virality from multiple disciplinary perspectives. Research in communication and marketing sciences has highlighted that factors such as emotional tone, novelty, and timing significantly affect audience engagement and sharing likelihood (3). However, earlier works were limited to static data and retrospective analysis, focusing primarily on post-hoc evaluations rather than predictive forecasting. Data mining models traditionally utilized techniques such as clustering, regression, and sentiment analysis to interpret existing trends, but these methods often fell short in adapting to the continuous evolution of social media algorithms and user behavior. Moreover, prior investigations tended to analyze quantitative engagement metrics without considering psychological, social, and cultural dimensions that strongly influence digital interactions. Consequently, the predictive accuracy of such models remained constrained, leaving a gap between computational analytics and real-world behavioral interpretation

Further advances in artificial intelligence, particularly in deep learning and neural networks, have provided new pathways for improving prediction accuracy. Studies employing these techniques have demonstrated promising results in modeling user preferences and content popularity (4). Yet, despite technological progress, a comprehensive understanding of virality that integrates computational, social, and psychological aspects remains limited. Many predictive systems are still designed around isolated datasets, neglecting contextual and emotional factors that drive audience decisions. There also exists a shortage of models capable of real-time adaptation—an essential feature in the rapidly changing digital landscape where trends emerge and decline within hours. This shortcoming highlights the need for holistic, adaptive frameworks that can operate across diverse social media environments and align with the societal goals envisioned under the Society 5.0 framework. To bridge these gaps, this study focuses on developing a predictive analytics model aimed at understanding and forecasting social media content virality. The proposed model serves as both a theoretical and practical contribution by integrating real-time data mining, engagement metrics, and association rule analysis into a unified system. Its core objective is to provide actionable insights for improving the performance of digital content and enhancing audience engagement. Recent studies have demonstrated the potential of such integrative approaches in strengthening business capabilities and enabling data-driven promotional strategies. Building on these findings,

this study extends the application of predictive analytics into the domain of social media virality forecasting, emphasizing interdisciplinary collaboration between computer science, behavioral studies, and communication theory. The incorporation of artificial intelligence allows the simulation of real-world promotional scenarios and content generation processes similar to systems such as DeepTweets and PixelCNN (5)

In addition to the computational framework, this research acknowledges the human dimensions of virality. Content dissemination on social media is not merely a function of algorithmic recommendation but also a reflection of collective behavior, cultural identity, and psychological resonance. By linking data-driven modeling with social science perspectives, the study situates predictive analytics within a broader understanding of societal interaction. This approach enriches the theoretical foundation of social media research by recognizing that virality is shaped by both measurable engagement parameters and intangible emotional triggers.

The overarching aim of this study is to construct a reliable and adaptive predictive model that can identify patterns leading to viral content dissemination. The specific objectives include: identifying key variables that determine the virality of digital content; designing a model that integrates computational, social, and psychological determinants; and validating the model through real-time data analysis and experimental evaluation. The novelty of the present research lies in its dual emphasis on technological and social factors, providing a multidimensional framework for predicting virality. By embedding predictive mechanisms into the analytical process, this model enables real-time decision-making that supports more strategic and human-centered content creation.

Ultimately, this introduction establishes the conceptual foundation for a detailed exploration of how data mining has evolved into predictive analytics within the context of Society 5.0. It also underscores how predictive technologies can empower organizations, marketers, and researchers to anticipate trends and design content that resonates with target audiences. The integration of predictive analytics into social media strategies not only enhances engagement but also contributes to achieving the human-centric ideals of Society 5.0. Through this research, predictive analytics is positioned not merely as a computational tool but as a transformative mechanism for understanding social behavior and advancing digital inclusivity.

Methodology

This study employs a qualitative modeling approach, utilizing both theoretical and empirical methods to construct and validate an abstract model of a real-world phenomenon, specifically focusing on the virality of social media content. The novelty of this study is the forecasting framework, which combines real-time scraping, engagement metrics, and association-rule mining to predict content virality. We aim to explore the causal relationships between various factors influencing social media engagement and their effects on content virality, distinguishing between dependent and independent variables within our model framework.

In this study, we identify a set of virality variables that represent measurable features of social media content and user interactions. These variables include engagement indicators such as likes, comments, shares, favorites, hashtag usage, and audio reuse, which are commonly used by platforms to determine visibility and recommendation. Each of these elements provides a quantifiable dimension of how users interact with content, making them suitable for inclusion in predictive models of virality.

The operationalization process involves three components. First, engagement metrics are directly extracted from social media posts to capture immediate user interaction. Second, association rules are applied to identify co-occurrence patterns among features such as hashtags and audio, providing interpretable insights into the combinations that contribute to viral outcomes. Third, real-time scraping is employed to ensure that the dataset remains up-to-date and reflective of current platform dynamics, thereby increasing the relevance and predictive power of the model.

Our research methodology integrates traditional scientific methods with modern computational techniques to accurately represent and analyze these complex systems (6). Through systematic observation, precise measurements, and iterative testing, we will adjust our hypotheses based on

experimental outcomes, ensuring the robustness and applicability of our model (7). The model's effectiveness and validity will be tested using a combination of data sources and experimental setups, which align with contemporary research practices in computational social science (8). Addressing experimental design, reproducibility, result evaluation, and ethical considerations is crucial to the integrity and success of research in pervasive computing and related fields (9). Figure 2 shows the systematic measurement process used in this study.

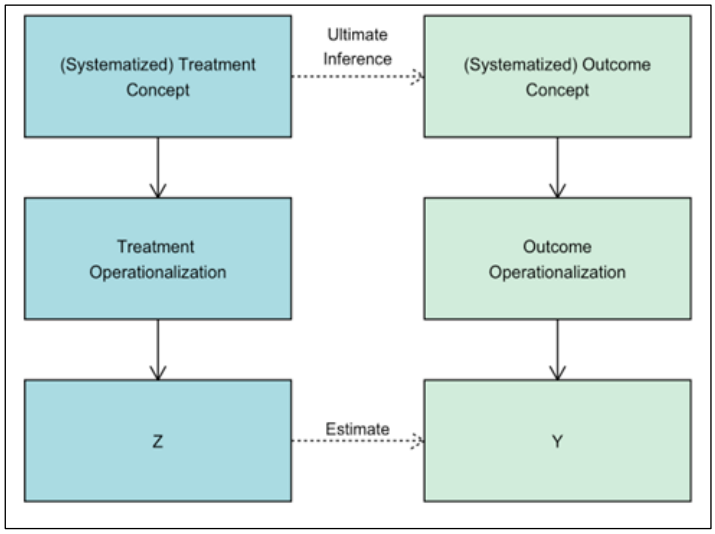

Figure 2: Measurement Process to Validate Data in Research (10)

To implement our research variables, we will implement a three-phase data collection strategy shown in figure 3. Initially, structured interviews will be conducted to gather qualitative data directly from users, which aids in capturing nuanced insights into user engagement and content interaction behaviors. Subsequently, a comprehensive literature review will be conducted to align our study with existing theories and findings in the field, thus grounding our research in a solid theoretical framework (11). Observational methods will also be utilized, where real-time data collection through social media platforms will provide empirical evidence to support or refute our theoretical constructions. This approach allows for the direct observation of interactions and engagements, offering a clear view of user behavior in naturalistic settings.

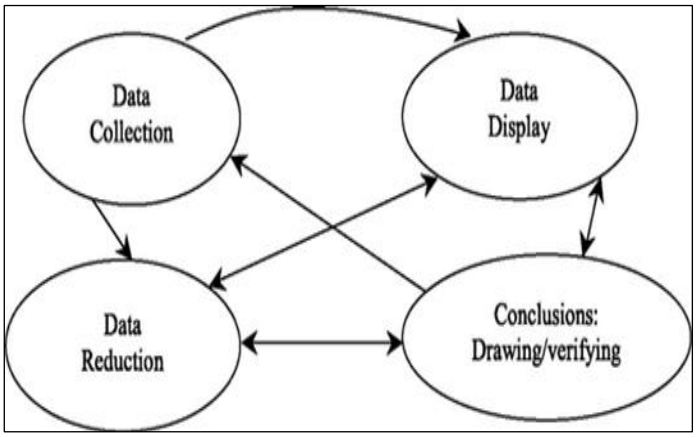

Figure 3: Data Processing Method

Data processing will involve three key stages: data reduction, data display, and conclusion verification. Initially, data reduction will simplify the complex data sets to manageable levels, focusing on significant variables that influence virality. This stage is crucial for identifying key patterns that will inform the subsequent phases of analysis. The processed data will then be organized and displayed in a structured database format, facilitating easy access and manipulation during

analysis. This database will contain detailed records of engagements and interactions, structured according to the predefined dimensions of our study.

The final stage involves drawing conclusions from the processed data. This phase will test the hypotheses by comparing the observed data against expected outcomes based on our predictive model (12). Verification through statistical and computational methods will ensure the reliability of our conclusions, ultimately reinforcing the validity of our research findings.

Formal measurement is also done by hypothesizing a treatment, Z_i and changing the preferences to democratic norms, v_i . The process is needed to validate the data and to estimate the causal effect of the research (13). The calculation for data validation is measured using object power which in our case is the virality. We will be using engagements of the treated object as our measurement foundation and other parameters of social media account experiments as comparison to validate our measurement results (14).

Through this comprehensive methodological framework, our study aims to advance the understanding of predictive analytics in social media settings, offering insights that could significantly benefit strategic content planning and digital marketing initiatives.

Results

Our model operates through a systematic process of data collection and experimentation with social media analytics. We collect raw data from social platforms, which are then transformed into discrete variables essential for analyzing content quality and engagement. This transformation allows us to dissect the dynamics of social media engagements—likes, shares, comments, and user interactions—which are critical drivers of virality. By quantitatively assessing these variables, we gain deeper insights into the mechanics of content spread and engagement, enabling the development of targeted strategies to enhance virality.

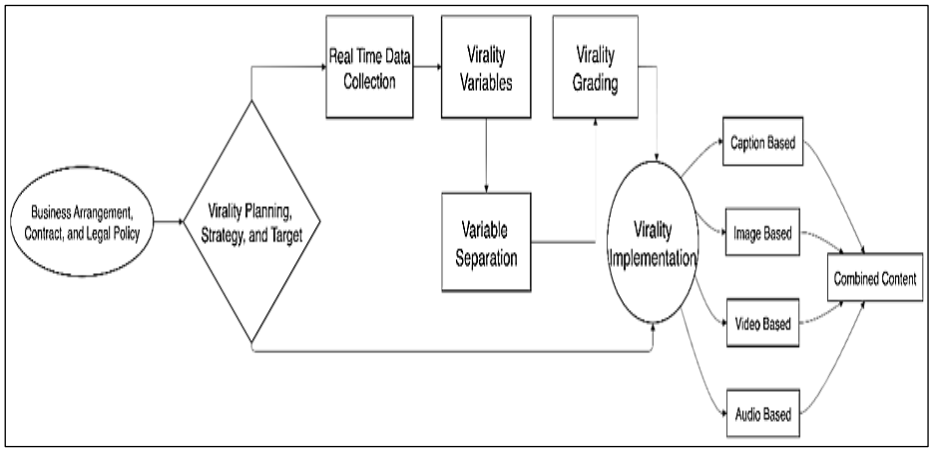

Figure 4: Proposed Model Workflow

The workflow of our model is detailed in figure 4 and begins with a real-time data collection phase. During this phase, we meticulously gather data pertaining to virality indicators. Subsequent to collection, we engage in a rigorous process to measure and analyze virality, using both traditional statistical methods and advanced machine learning techniques. This virality measurement process evaluates the potential of content to become viral by considering factors such as engagement rates, dissemination speed, and the demographic characteristics of the engaging audience.

Incorporating a market basket analysis approach, our machine learning algorithm scrutinizes large datasets to discern patterns and correlations among the collected variables. This analysis helps in understanding how different engagement metrics interact to influence content's viral potential. By identifying which combinations of engagement factors commonly lead to increased virality, the model can predict which new pieces of content are likely to achieve widespread popularity.

Our analytical process is iterative and responsive; we continuously monitor the performance of our virality models and refine our strategies based on

empirical evidence. This adaptive approach ensures that our predictions remain relevant and accurate, allowing for real-time adjustments to our virality enhancement techniques. This data-driven approach not only helps in crafting compelling content but also in making informed marketing decisions that can significantly amplify the reach and impact of social media campaigns (15).

The final stages of our model focus on strategic application and implementation. We utilize the insights and data derived from our analysis to inform and guide promotional strategies, aiming to optimize content for maximum engagement and virality. Our approach should explicitly incorporate psychological drivers because high-arousal emotions (for example awe, anger, or amusement) are empirically strong predictors of sharing and therefore improve predictive performance when included as features (16).

Through this comprehensive model, we strive to provide a robust framework for understanding and harnessing the factors that drive social media content virality, ultimately aiding businesses in effectively engaging with their audiences and achieving their promotional goals.

Business Planning Phase

The first phase of our proposed model contains business planning to combine the business' target, category, environmental influence, period, and media which will be used as inputs. This also acts as a legal standing between the business as the third party and us as the executive using our proposed model.

During the business planning phase, the inputs mentioned above will be analyzed and used to create a comprehensive plan for the business. This plan will outline the specific goals and objectives of the business, as well as the strategies and tactics that will be used to achieve them. The plan will also take into account any external factors that may impact the business, such as competition and market trends. This phase is crucial for ensuring that the business is on the right track and that all stakeholders are aligned on the direction of the business. Additionally, it serves as a legally binding document between the business and the executive team implementing the proposed model.

After the plan has been created and reviewed, the next step is to execute it. This typically involves creating a project team, assigning roles and responsibilities, and setting deadlines for key milestones which will be combined in our proposed model. The project team will then work together to implement the strategies and tactics outlined in the plan, using the inputs provided during the business planning phase as a guide. This may involve creating marketing campaigns or making changes to existing strategies to gain viralities.

As the project progresses, the project team will monitor progress and performance and adjust as needed. This may involve revising the plan or changing course if certain strategies are not working as expected. The business planning phase will also involve regular reviews and evaluations to ensure that the proposed model is on track to achieve its goals and objectives.

As our findings we can confirm that the business planning phase is a critical step in the success of the business, as it lays the foundation for the entire project and ensures that everyone is working towards the same goals.

Business Arrangements

The initial phase of our proposed model involves a detailed business planning process as shown in figure 5. This phase integrates various inputs such as business targets, category, environmental factors, timeframes, and media choices. These elements not only guide the planning but also establish a legal framework for the interactions between the business, as the client, and our team, as the solution providers.

During this crucial phase, we meticulously analyze the inputs provided to develop a strategic business plan. This plan outlines the specific goals and objectives of the business and devises strategies and tactics aimed at achieving these goals. It considers external factors like market competition and industry trends, which are vital for adapting the strategy to real-world challenges.

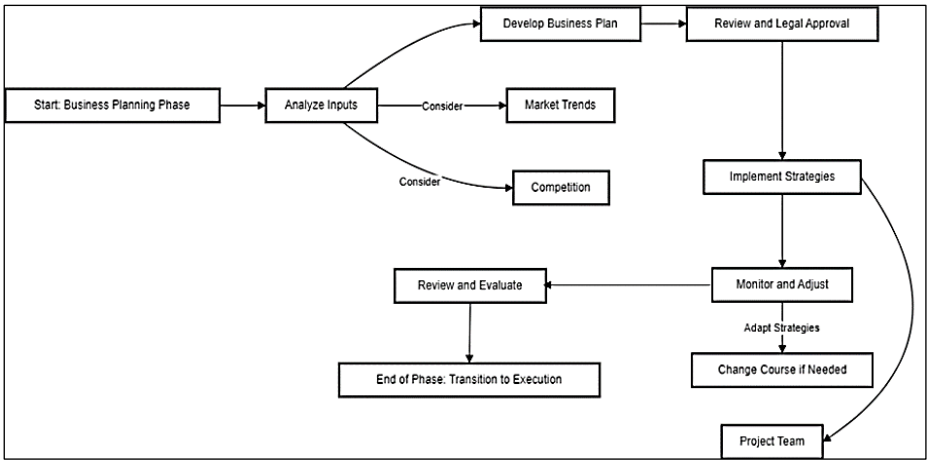

Figure 5: Business Planning Phase

Once the business plan is reviewed and legally approved, we proceed to execute it. Execution involves assembling a project team, assigning clear roles and responsibilities, and setting deadlines for critical milestones, all guided by the strategic framework established in the planning phase. The project team collaborates to enact the outlined strategies, potentially launching marketing campaigns or tweaking existing strategies to optimize virality. Throughout the project, our team continuously monitors progress and performance, making necessary adjustments to stay aligned with the objectives. Regular reviews and evaluations are conducted to ensure the efficacy and alignment of the business plan with the project goals. The business planning phase is fundamental to the project's success, laying a robust foundation for all subsequent activities and ensuring cohesive efforts towards the shared objectives.

Contractual Framework in Virality Technology Implementation

A contract in the context of deploying a virality technology model represents a legally binding agreement that delineates the terms and conditions under which the technology will be utilized. This includes stipulations on how user data will be collected and managed, restrictions on the use of the technology, and financial terms concerning the usage of the technology. Additionally, the contract specifies the legal responsibilities and liabilities for any potential issues that might arise from using technology. Essentially, the contract is instrumental in clearly defining the rights and duties of all parties involved in the technology's deployment and operation.

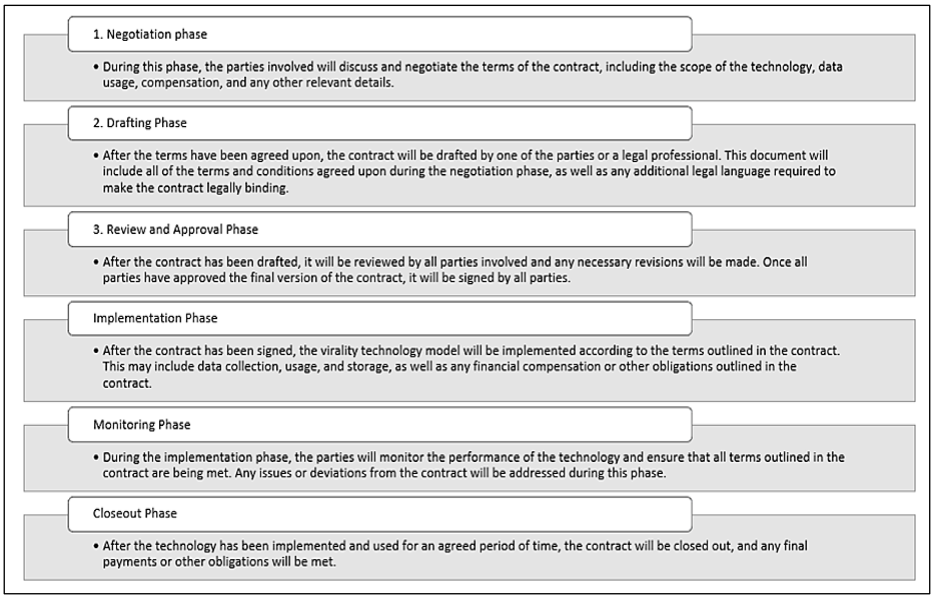

Figure 6: Model Implementation Contract Phase

The contractual process is a fundamental component of implementing a virality technology model, ensuring that all parties' rights and obligations are explicitly articulated. This process encompasses several critical stages shown in figure 6.

Each phase is crucial to guarantee that technology implementation is successful and that the interests of all stakeholders are safeguarded. It is vital that the terms of the contract are clearly understood and agreed upon by all parties before its execution to prevent any misunderstandings or legal complications during the implementation of the virality technology model.

Legal Policy Framework for Virality Technology Model

The legal policy for a virality technology model as shown in figure 7 is comprehensive, aimed at ensuring ethical usage while safeguarding user privacy and data protection. It encompasses guidelines that govern the responsible usage of technology, addressing critical issues such as intellectual property rights and potential liabilities resulting from adverse effects of technology use. Moreover, the policy includes stringent provisions for regulatory compliance and oversight to ensure operations remain within legal and ethical boundaries.

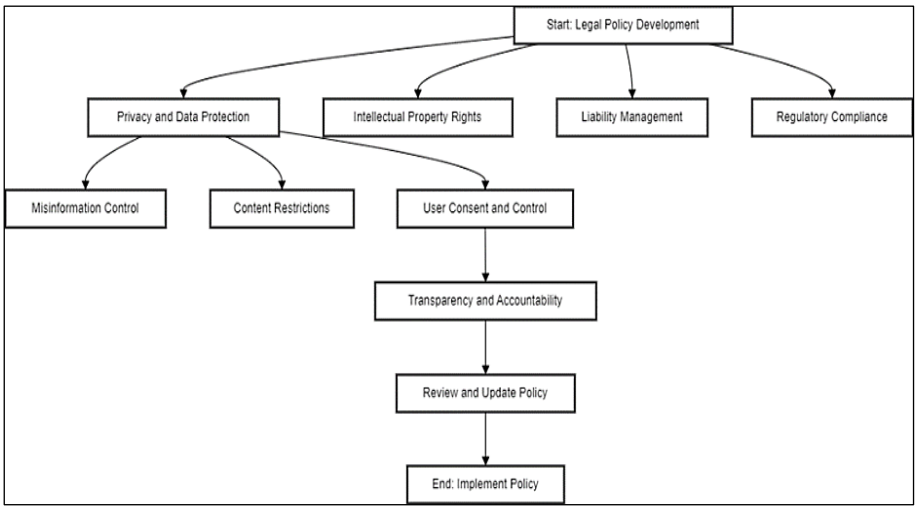

Figure 7: Legal Policy Framework

To address the proliferation of misinformation and disinformation, the policy mandates procedures to verify the accuracy of the information circulated through the platform, coupled with measures to curb the spread of misleading content. It also tackles the potential misuse of the technology for harmful purposes, such as spreading propaganda, by restricting certain types of content and employing monitoring systems to prevent misuse.

The aspect of user consent is critical; the policy must explicitly outline the types of data collected, their usage, and offer users the ability to opt-out or manage their data preferences.

Furthermore, the policy emphasizes transparency and accountability. This includes regular updates on technology usage and the establishment of an effective complaint and redress mechanism for users to report grievances or rights violations.

Virality Technology Preparation Phase

The preparation phase for our proposed virality technology model encompasses comprehensive planning, strategic development, and target setting. This phase includes three crucial components: virality planning, virality strategy, and virality target.

Virality planning as shown in figure 8 involves identifying the specific goals and objectives of the model, as well as determining the target audience and the platforms on which the campaign will be promoted. Key metrics will also be identified to measure the success of the campaign. These include metrics such as the number of views, shares, or conversions that the virality project aims to achieve.

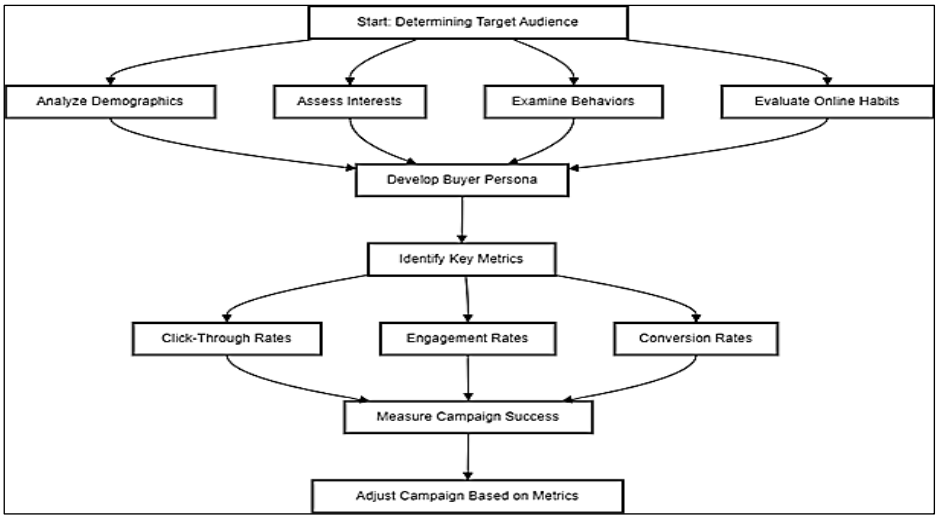

Figure 8: Virality Planning

When determining the target audience, the proposed model considers demographics, interests, behaviors, and online habits prevalent on social media. This information is used to develop a buyer persona or a profile of the ideal customer for the campaign. The model will then operate according to the key metrics identified in the previous phase to measure campaign success. Metrics like click-through rates, engagement rates, and conversion rates provide insights into campaign performance and allow for necessary adjustments to improve results.

In summary, virality planning is essential for the preparation phase of a viral campaign. By carefully planning and executing a campaign, businesses can significantly increase the chances of creating a viral sensation and achieving their desired outcomes.

Virality strategy involves developing a detailed plan for how the project will be executed. This step includes identifying the type of content to be used, the messaging strategy, and the tactics to encourage sharing of the content. Based on our analysis, certain types of content, such as well-known objects or persons, text-to-speech audio, trending audio and popular topics or hashtags, are more likely to resonate with the audience. Further research is needed to understand the psychological impact of these content types on the audience. Figure 9 shows a sample of well-known object.

Figure 9: Well-known Object Sample Implementation in Virality Technology Analysis

Our video experiments using well-known or popular objects have shown significant increases in virality and engagement. For instance, using the University of Indonesia logo, UKSW University, Indomaret supermarket, She-Hulk videos, and talents from popular people on campus resulted in positive outcomes. These experiments demonstrated not only an increase in virality but also a boost in engagement and video views retention.

Incorporating recognizable and relatable elements into our videos allowed us to tap into existing audiences and generate more interest. Featuring the University of Indonesia logo reached a large audience of students and alumni, while including She-Hulk videos attracted fans of the Marvel Cinematic Universe. Handpicked talents from popular campus figures brought authenticity and credibility, fostering a sense of community and connection with viewers. This approach led to greater success in terms of virality, engagement, and retention.

Virality target refers to the specific group of people that the campaign aims to reach. This includes demographics such as age, gender, location, and interests, as well as behaviors like online activity and purchasing habits. Setting a clear target audience ensures that the campaign reaches the right people, increasing its likelihood of success.

By identifying a specific group of people, we ensured that our content reached the appropriate audience, enhancing its relevance and engagement. For example, targeting students and alumni of the University of Indonesia for videos featuring the institution's logo, and fans of She-Hulk and the Marvel Cinematic Universe for related content. Understanding the behavior and purchasing habits of our target audience, such as frequent shoppers at Indomaret supermarket, allowed us to create content that resonated with them, increasing the chances of success.

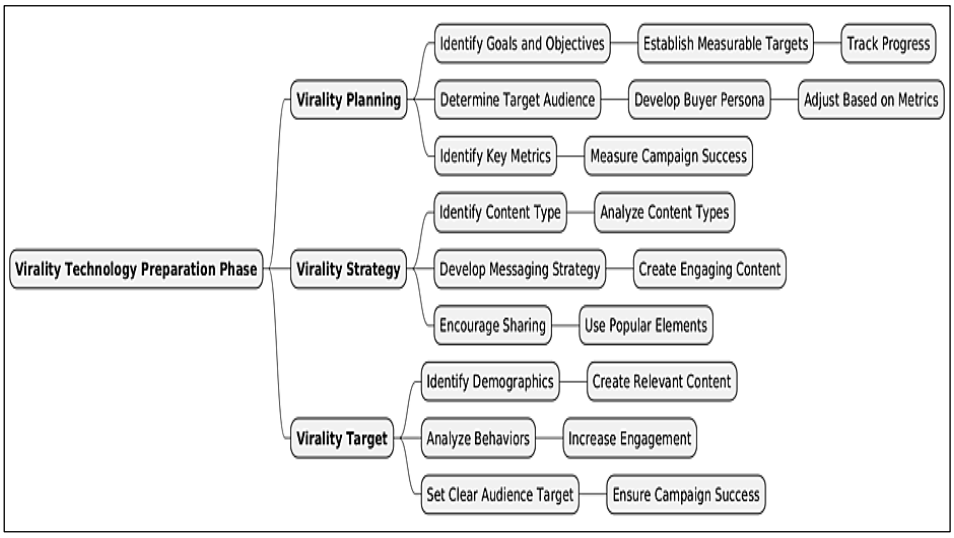

Figure 10: Virality Technology Preparation Phase

The diagram on figure 10 illustrates the structured approach to preparing the virality technology model, highlighting the phases of planning, strategizing, and targeting, each with its specific activities and goals. By systematically planning, strategizing, and targeting, businesses can optimize their virality campaigns to reach and engage their desired audiences effectively, leading to increased success in achieving viral content.

Real-time Data Collection

In our proposed model, we identify and collect virality variables using parallel mode in a real-time activity to ensure the dataset is fresh and increase the possibility of our next content to be viral. Once the virality variables are collected, they are then processed and analysed to identify patterns and trends. Our reliance on streaming and sampled platform endpoints implies potential sampling bias that must be acknowledged and mitigated in methods and interpretation (17). These insights are then used to inform the creation of new content, with the goal of increasing its chances of going viral. Additionally, the model can also be used to optimize the distribution and promotion of existing content, to further boost its virality. These activities include three main pillars: conventional, data mining, and data scrapping technique as shown in Figure 11.

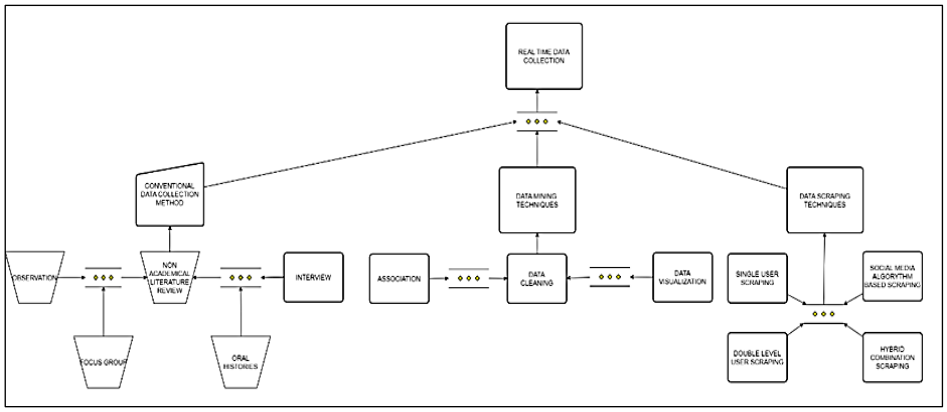

Figure 11: Real-time Data Collection Structure

The model can also be fine-tuned over time by constantly updating the dataset with new data and adjusting the algorithm based on the performance of previously created content. This allows for continuous improvement and the ability to adapt to changing trends in the online environment. Virality prediction benefits from online/adaptive modeling that update parameters as new events arrive and can explicitly model timing and external promotion inputs (18). Furthermore, the model can be used to predict the virality of content before it's even created, giving content creators an idea of what types of content are more likely to be successful. Our proposed model aims to provide a comprehensive solution for identifying and creating viral content, to help businesses and individuals increase their online visibility and reach by predicting virality and use it as promotional content.

Conventional Data Collection Method

We try to explore three conventional methods and add it to our model for collecting data in the context of virality technology: manual observation of social media, non-academic literature review, and social interviews.

Manual observation in social media involves manually reviewing posts, comments, and other user-generated content on various social media platforms to gather information about how users interact with and share viral content. Manual observation remains valuable to capture contextual signals, discourse, and motives that automated scrapers miss, but it must be presented as a complementary, non-representative qualitative technique (19). This method is used to gather different aspects of virality such as the types of content that are most likely to go viral, the characteristics of users who share viral content, and the factors that influence the spread of viral content manually. One advantage of manual observation in social media is that it allows us to gather viral content in its natural context and to gather detailed information about how users interact with and spread viral content which is hard for data scrapping technique to gather. However, manual observation can also be time-consuming, and it may be difficult to obtain a representative sample of social media posts or users which is why we add automated data collection using data scrapping techniques to further increase the effectiveness of our model.

Non-academic literature review involves reviewing books, articles, and other publications that are not peer-reviewed academic papers to gather information about the history, theory, and current practices of virality technology.

This is a useful method for collecting data on history, theory, and current practices of virality technology which circulate around the internet. The data collected through this method can include information on the key concepts, theories, and models that have been developed to understand and predict virality, as well as practical examples of how virality technology is being used in different industries and contexts. This data can be collected from a variety of sources, such as books, articles, blog posts, and online forums, and can be analyzed to identify patterns, trends, and insights that can inform the development of virality technology models. Additionally, non-academic literature review can also include conducting interviews with experts in the field to gather their perspectives and insights on virality technology. Surveying grey literature and practitioner sources complements peer-reviewed studies by revealing up-to-date platform practices and emergent

strategies, but findings should be triangulated with empirical trace data to avoid anecdotal bias (20).

Social interviews involve conducting interviews with individuals who have experience creating, sharing, or studying viral content to gather first-hand accounts and insights about the virality process. Structured social interviews provide first-hand explanations of posting strategies and motivations that help interpret quantitative patterns and improve feature design for predictive models (21). These can provide valuable information about the factors that contribute to the success of viral content, such as the type of content, the audience it is aimed at, and the platforms on which it is shared. Additionally, it can also provide insight into the motivations and strategies of individuals who create viral content, as well as the ways in which viral content can be used to achieve specific goals, such as increasing brand awareness or driving engagement. Overall, social interviews can be a powerful tool for understanding the dynamics of viral content and how it can be leveraged for various purposes.

Association, Data Cleaning and Data Visualization are the key techniques we used for our virality technology model to analyze and understand the social media data. The model can predict which posts are likely to become viral and help users to improve their content strategy. This method is done simultaneously and correspondingly with other activities in the data mining activity. Below we further discuss the workflow of our data mining techniques we used in our model.

Association rules are used to identify patterns in social media data that can help predict which posts are likely to become viral. For our model, we use market basket analysis to capture real-time data which includes captions, engagement totals, number of used hashtag and audio. We then divide the results based on the confidence grade of the social media variables. Market-basket and association-rule mining are appropriate here to discover frequently co-occurring feature-sets (captions, hashtags, audio) that correlate with high virality power and provide interpretable rules for the virality engine (22). Table 1 shows the rules we set for both virality prediction and confidence grade of the social media variables.

Table 1: Proposed Model Association Rule

NO.	RULES
1.	User Content → Viral (V_i), where $\{e_1, e_2, \dots, e_n\}$is >= Max Total of User's e
2.	Engagement (e)→ Valid, IF Single Engagement Variable Total (e_t**)** AND Average Total Engagement (e_μ**)**, TRUE
3.	Virality Grading (V_g)→ Valid, IF Positive Grades (P_g) AND Negative Grades (N_g), TRUE

In order to validate virality using our proposed model, we need to identify user content virality by analyzing the engagement of all their social media content to determine the max total of engagement which will be the base of calculation for our virality measurement. We also need to consider and count availability of all engagements so that we can scrap using our scrapping technique.

We also need to state a rule in terms of grading and divide it into positive grades to show effective videos which we can use as our dataset later and vice versa. Depending on the data we scrap; we can divide grading into different level of categories. In our experiments, we divided it into very bad, medium bad, bad, normal, good, very good, and best.

This allows us to have a clear understanding of the quality of the videos in our dataset and ensure that we are only using high-quality videos for our experiments. Additionally, having a clear grading system in place will make it easier to analyze and interpret the results of our experiments. It also provides a consistent method for evaluating videos, which is crucial for ensuring the validity and reliability of our findings. Furthermore, it will allow us to make comparisons between different videos in our dataset and identify any trends or patterns that may be useful in informing our research. Overall, a well-defined grading system is essential for ensuring the accuracy and credibility of our experiments and findings.

Data Cleaning

Systematic data cleaning — deduplication, missing-value treatment, canonicalization of tags/hashtags, emoji normalization and outlier handling — is essential to avoid downstream biases and to ensure reproducible feature extraction (23). It is an important step in the

process of scraping social media variables such as engagement numbers, captions, hashtags, and links. The first step in our proposed model is to remove any duplicate entries in the data to ensure that it is unique and free of unnecessary repetition, this process is needed only for single user media or influencer scrapping target. Next, missing values must be handled by either removing the rows that contain them or inputting them with suitable values. The data must also be formatted consistently, with engagement numbers in numerical format, captions in string format, and hashtags and links in list format. Unwanted characters, such as emojis, should be removed to make the data more manageable, and variations in the way data is entered, such as hashtags written with or without the "#" symbol, should be normalized. Finally, common words known as stop words which don't add much meaning in the analysis should be removed. These steps help to ensure that the data is clean, consistent, and ready for analysis. Figure 12 shows the proposed model data cleaning process.

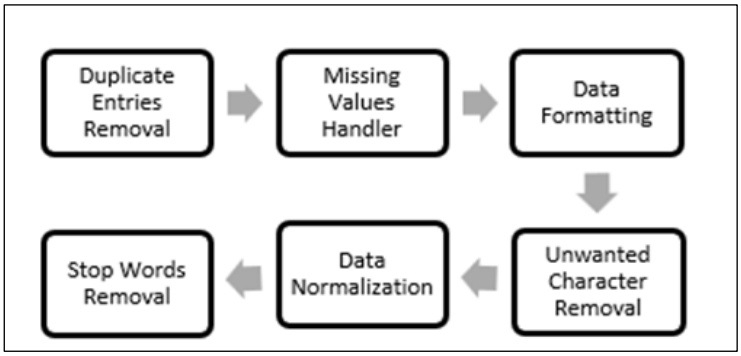

Figure 12: Data Cleaning Process in Our Proposed Model

Additionally, we also check for outliners in the engagement numbers and remove them if they do not align with the rest of the data. We also check for any irrelevant hashtags or links that may not be relevant to the specific analysis and remove them. The next step is to check for any irrelevant captions and remove them. In this way, we can make sure that the data is accurate and relevant to the analysis. Once the data is cleaned, it can be transformed and consolidated into a format that is suitable for further analysis such as creating a data frame, and then saved for future use. This is a crucial step in the process of scraping social media variables as it helps to ensure that the data is accurate and relevant, and ready for further analysis.

Data Visualization

The cleaned data is then analyzed using various techniques which correspond to the data we scrap such as descriptive statistics, clustering, and machine learning algorithms to identify patterns and trends. These techniques help to identify the factors that contribute to the virality of a post, such as the use of specific hashtags or topics.

Figure 13: Word Cloud Visualization of Scrapped Hashtag Data on August, 19th 2022

The output of this process is presented in the form of visualizations such as charts, graphs, and heat maps. Visual analytics (heatmaps, time-series dashboards, word clouds and retention curves) are indispensable for exploratory analysis and for communicating patterns, but visualizations must follow good design practices to avoid misleading interpretations (24). These visualizations provide insights into the data and can be used to suggest the best audio, caption, topic, and hashtag to use for a social media post to increase its chances of going viral. Figure 13 shows the word cloud visualization of the scrapped data used in this study. Suggested audio results can simply be visualized by the count it used based on the scraped data. It should be relevant enough to use for the next predicted social media post virality.

Data Scrapping Techniques

As we finished the conventional data collection and data mining in parallels using our model. We also add data scrapping activities to our parallel activity to collect data in real-time. Post-API platform constraints and legal/ethical considerations require that scraping workflows record provenance, respect terms of service, remove personal identifiers when necessary, and document institutional approvals where applicable (25). Our data scrapping techniques which focus on harvesting data from social media consist of: single user scrapping which we focused in scrapping potential viral content from influencer or other known media, double level scrapping where we first focus on scrapping media links due to the scrapping restriction, social media algorithms-based scrapping where we adapt our scrapping tool to gather data based on the structure and algorithms of the social media, and also combined scrapping techniques from above to improve the result of our scrapping activities.

Single User Scrapping

The goal of scraping is to gather information on a specific user's social media activity as our first phase method to gather sample and media links which we will use later in our model to produce viral social media post. This information can also be used for various purposes, including market research, content analysis, and insight reports to the stakeholder. In our proposed model, single user scraping method as shown in figure 14 is focused on scraping content links.

Scraping content links from a single user's social media account involves using a web scraping tool or software to extract data from the platform. This tool or software will visit the user's profile, collect the links to their content, and store the data in a structured format using the previous method discussed. The data collected can then be analyzed to gather insights into the user's content and activity and used to produce viral social media content later. Once the data is collected, the model will analyze the data to gather insights. Aside from media links, we also gather the frequency of content posts, the type of content posted, and the basic engagements insight such as total views, number of comments, likes, etc.

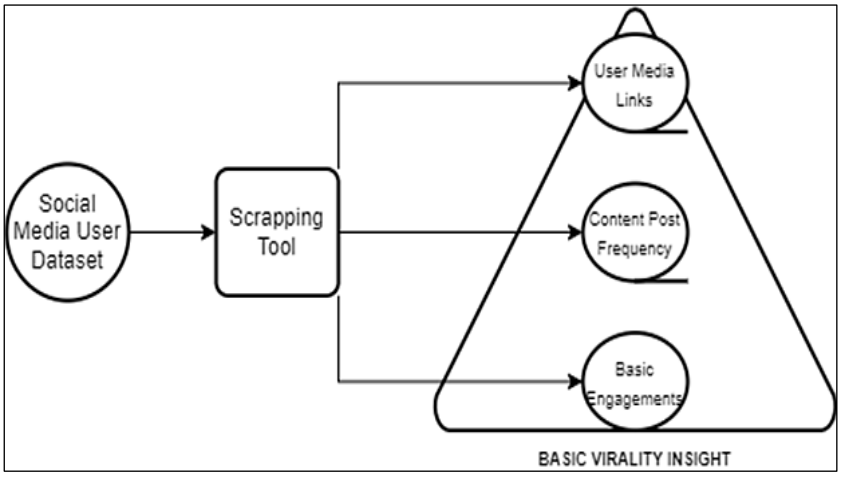

Figure 14: Single User Scrapping Method in Our Proposed Model

However, this method only acts as a preparation activity and should be further combined with double level user scrapping, social media algorithms-based scrapping, and hybrid combination scrapping which we used during the experiment.

Double Level User Scrapping

Continuing the scrapping activity previously, we then take the media links from dataset to be put into second stage scrapping we called Double Level User Scrapping as shown in figure 15. The word double means that we dive into the second scrapping activity using the data from the first stage scrapping which is the single level user scrapping. In this method, we took scrapped content links from our targeted user social media to begin scraping necessary data that will be used for our proposed model analyzing process to produce virality content. Depending on the social media itself, engagements data that we extract would vary.

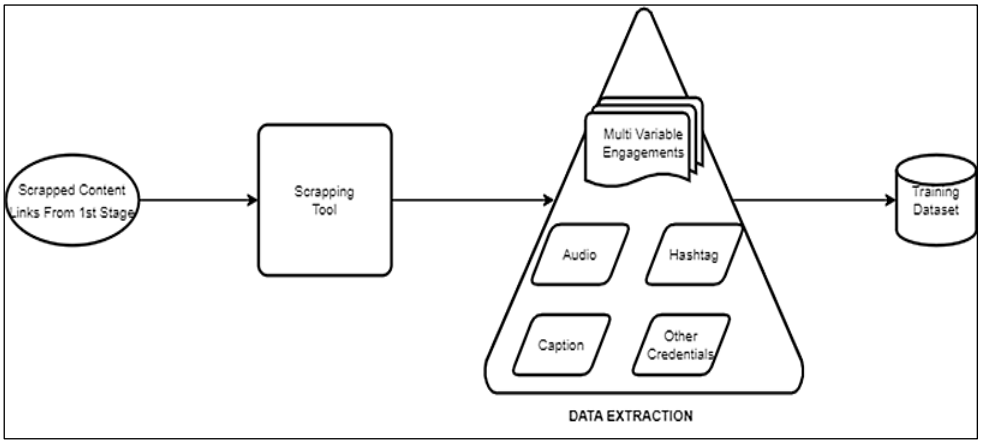

Figure 15: Double Level Scrapping Method in Our Proposed Model

For example, Tiktok engagements data would contain likes, comments, shares, and favourites. Aside from engagements, we also extract all possible data to be used which could contain audio, captions, hashtag, and other credentials such as: tagged user, user credentials, and date. The extracted data will then be stored as training dataset for the next phase together with other scrapped data we extract from conventional data collection and analyze it with our data mining techniques that we use in our proposed model.

Social Media Algorithm-Based Scrapping

Social media algorithm-based scraping involves using an automated tool to collect data about popular accounts and influencers on social media platforms such as Twitter, Instagram, or YouTube. As the structure of the social media is different from each other, we need to design different scrapping workflow for each social media and update it if necessary, since the structure of social media itself changes over time. Figure 16 shows scraping workflow used in Tiktok For You page.

Figure 16: Sample Workflow for Tiktok for You Page Scrapping Workflow

In this scrapping activity, we focused on extracting user account links which will be directly used as foundation link to engage single user scrapping activity. This information is then added to the dataset and analyzed to determine the influence and reach of these accounts. The factor analyzed includes the number of followers, engagement rate, and content relevance.

This information can also be used to study trends and patterns, target advertising, and improve the overall performance of our proposed model. The analysis of the dataset can also be used to identify key influencers and their role in shaping public opinion. Additionally, the results of this scrapping activity can be used to monitor changes in the user's account activity, as well as to track the growth and decline of their online presence over time.

Hybrid Combination Scrapping

In this scrapping activity, we combine the three scrapping activities which will run in parallel mode since we design the proposed model to collect data in real-time in order to get the best results. This activity acts as a controller for the three scrapping activities which also determine the workflow of those main scrapping activities. Operationally, hybrid scraping (algorithm-aware crawling + single-user and second-level media expansion) improves coverage but must be balanced against rate limits and ethical safeguards; governance and logging facilitate reproducibility and audit (26). Figure 17 shows the workflow of hybrid combination scrapping activity.

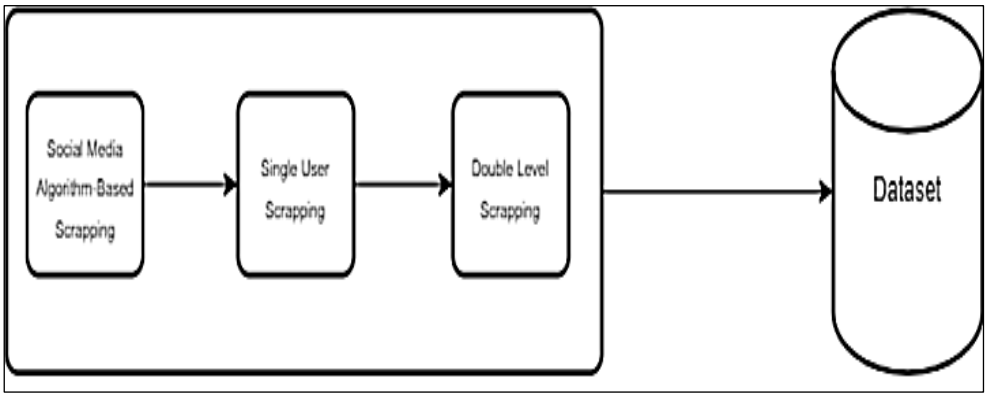

Figure 17: Hybrid Combination Scrapping Workflow

Since the data extraction depends on each scrap's result, we need to create a scrapping workflow which controls the flow of the data extraction between each activity. Differences in recommendation are major confounders in cross-platform comparisons and should be modelled or controlled for in any causal claim about intrinsic content contagiousness (27). The workflow begins with the social media algorithm-based scrapping which took the username and link from the platform. The user links are then used in the single user scrapping to provide media links and basic engagements data. The media links are then used to extract in-depth data and other engagements variables which then will be put in the dataset for our proposed model next phase.

Our model focuses on processing the virality variables by separating it and analyzing it. In our proposed model measurement process, we separate variables into individual engagements items (e) which are then used into calculations. The numbers of the e will depends on the platform of the data collection activity were executed. We classify the e into 4 items: likes engagement (e1), comments engagements (e2), share engagements (e3), and other object variables (e_n) which are different variable depending on the social media platform such as favorites, audio uses count, hashtags, topics, etc. Figure 18 shows the variable separation workflow.

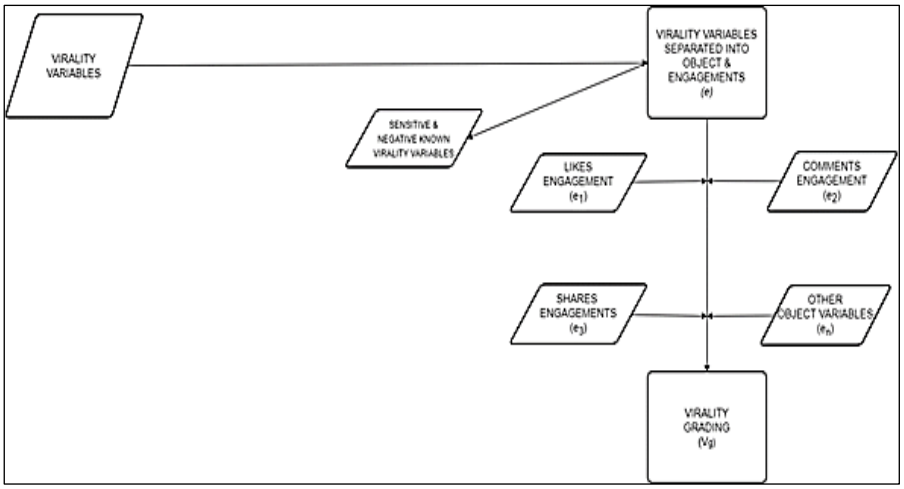

Figure 18: Virality Variable Separation Workflow

Virality Measurement

After the separation process, we conduct experiment with the e to determine virality while following our association rules to validate our findings. The separation process refers to the process of separating data by using data mining tools into various subsets to enhance the virality gain and its effectiveness as stated before in the hypothesis. Recent deep and graph-based predictive approaches demonstrate that content virality can be forecast with measurable accuracy before publication when structural and temporal features are available, supporting pre-publication optimization (28). In this case, the aim is to determine the virality of the variables while following the association data mining rules we use in the experiments. This is done by experimenting with variables to find a relationship between them and the outcome, the degree of virality. In our experiment, we create an equation with a "bottoms up" approach to solve.

We need to classify the variables e into separated numerical variables to increase the accuracy of our calculations to measure the virality of single social media content. By converting categorical variables into numerical values, we can then apply our experiment virality calculation and algorithms to make predictions about the virality of social media content in our proposed model. Accuracy and virality impact increases according to the number of e that we collect during the real-time data collection phase.

To measure virality, first we use the virality grade formula, we then separate virality grades into negative and positive grade by counting the number of grades (g_t) and multiply it with the divided grade value (n) then subtract it with average total of all collected variable (e_μ**)** to determine negative grades (N_g**)** or add e_μ to determine positive grades (P_g) as shown in the equation [1] and [2].

$$N_g = (g_t \times n) - e_\mu \qquad [1]$$

$$P_g = (g_t \times n) + e_\mu \qquad [2]$$

g_t **=** Divided grade value

n = Number of grades

e_μ **=** Average total of all collected engagement variable

Secondly, we divide the grade of the positive or negative grades. We came up with dividing formula using the number of dividing grades (g_n) as its core. We divide g_n with e_μ to show result based on how good the impact of the content based on its individual variable. The dividing formula provides a quantitative method for evaluating the impact and quality of content. By using this formula in our proposed model, we can improve the produced content and ensure that it has maximum impact on the target audience if combined with other variables. In our experiment, we divide the grade into 3 positive grades: Good, Very Good, and Best, and 3 negative grades: Very Bad, Medium Bad, and Bad. The formula is shown in equation [3].

$$V_g = e_\mu / g_n \qquad [3]$$

g_n **=** Number of dividing grades

V_g **=** Virality Grade

e_μ **=** Average total of all collected engagement variable

Thirdly, we validate e before determining virality to strengthen our result. In order to validate$e$, we came up with crude measurement by subtracting the total of an engagement variable with the

average total of all collected engagement variable (e_μ) in its group. By subtracting with average total, we can see the grade of the engagement and categorize it into positive or negative grade level. The validation of e formula is shown in equation [4].

$$e = e_t - e_\mu \qquad [4]$$

e = Engagement Variable such as: Number of Likes, Comments, Shares, etc.

e_t = Single Engagement Variable total

e_μ = Average total of all collected engagement variable

Lastly, virality is determined if the total of result is higher than the average engagements total after validating$e$. The equation calculates and determines virality V_i using social media engagements total as shown in equation [5].

$$V_i = e_1 + e_2 + \cdots + e_n \qquad [5]$$

e = Engagement Variable such as: Number of Likes, Comments, Shares, etc.

V_i = Viral if $V_i >$ =Max total of e_μ

Discussion

We conduct three experiments as a part of our virality implementation and use time-based variable to ensure the retention rate of our content, face/object variable in part of the retention rate, and other important variables such as trending hashtag, topics, and audio as a part of our content. Figure 19 shows the process of our virality implementation.

Our implementation consists of social media content in the form of reel videos using the output from previous phase. As a part of our experiment, we use Tiktok and Instagram platforms to test our virality measurements and variables and analyze the results. Table 2 shows the details of our first experiment.

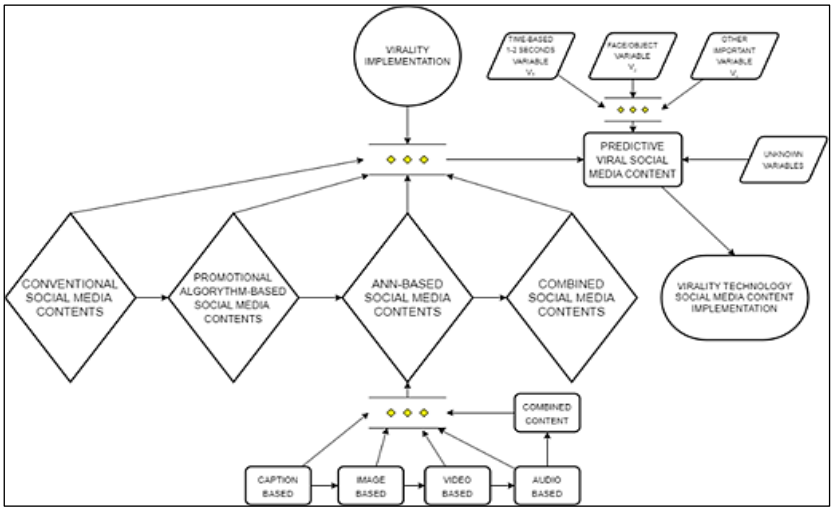

Figure 19: Virality Implementation Process

Table 2: First Experiments Result

No	Media ID	e_1	e_2	e_3	e_4	e_t	Power
1	7125241695268031771	755400	26200	461	2727	784788	3065.578
2	7127826243134819611	826300	5303	140	66	831809	1155.290
3	7128317294299712795	189800	6284	206	191	196481	249.341
4	7145448100423863578	1900000	69900	366	69	1970335	160.151
5	7112631468379131162	21600	1134	42	7	22783	149.888
6	7127074759782845723	83500	3872	153	48	87573	146.199
7	7117487925729791259	19700	538	58	2	20298	133.539
8	7124025223212567834	29700	300	18	2	30020	128.841
9	7130812987908164891	396800	44100	315	783	441998	100.000
10	7134189966959152410	450900	21000	445	869	473214	56.068
11	7128566932118277403	143300	1322	228	20	144870	34.575

12	7128748688918187291	127300	1614	242	18	129174	30.829
13	7136334060657446171	210900	23800	105	95	234900	20.930
14	7136217825890274587	213200	20400	78	108	233786	20.831
15	7138569642431237402	206500	7858	150	51	214559	17.440
16	7127815781777198362	10200	421	20	8	10649	14.790
17	7129811974199790875	56600	3612	80	5	60297	14.288
18	7125988552852147482	3675	175	10	2	3862	11.994
19	7141381585479257371	117100	8707	73	74	125954	10.238
20	7126352341749550362	2070	25	5	1	2101	6.465
21	7127447807870455067	3528	184	29	4	3745	5.852
22	7131329039045758235	16500	505	2	2	17009	3.817
23	7128351129087397147	2978	23	0	1	3002	3.810
24	7147665620094471451	40600	1742	7	1	42350	3.442
25	7146226760131169563	36400	3021	36	5	39462	3.208
26	7127648762616433947	1561	132	15	7	1715	2.680
27	7128239389263564059	1592	40	2	0	1634	2.074
28	7139913346115013914	18000	4744	130	6	22880	1.860
29	7136743336039189786	20300	707	5	5	21017	1.708
30	7137294542495124763	19400	952	68	11	20431	1.661
31	7135105908832488730	13100	72	3	2	13177	1.427
32	7134175594400599323	10100	1814	1	4	11919	1.412
33	7148577647151418650	14300	428	19	0	14747	1.199
34	7130537698909031706	4419	44	1	0	4464	1.029
35	7133900002178436379	8084	80	6	4	8174	0.993
36	7142048352870337819	9711	1388	2	2	11103	0.902
37	7134712725778795802	7393	209	7	9	7618	0.884
38	7143968266736749850	7420	112	2	1	7535	0.612
39	7132482074698681626	2534	166	1	3	2704	0.580
40	7139151854537035034	4652	646	18	8	5324	0.433
41	7144666463746477339	4901	160	0	2	5063	0.412
42	7131375385354226971	1757	48	2	1	1808	0.406
43	7130444966668569882	1116	20	3	1	1140	0.263
44	7132687982242450714	1590	16	1	4	1611	0.260
45	7137931969966443803	2417	100	1	0	2518	0.205
46	7137499430973984026	2301	67	9	1	2378	0.193
47	7136389676331453722	1718	90	4	1	1813	0.162
48	7132808356909174043	691	6	8	1	706	0.114
49	7149049760970788123	955	15	0	1	971	0.079

Our first experiment consists of 49 simulations of reels video which we share throughout Tiktok account: PeablePebble which focuses on education-based topic. In this experiment we try to validate the variable we use and measure the power of its virality. To determine the power of virality, we use the followers count as base measurement and divide it with engagement total (e_t) which should have a minimum of 1 point to be determined successful. Based on the experiment, 34 out of 49 simulations show positive results where the highest score count is 3065.578. This means that the content was able to reach an audience more than 3065 times greater than the

follower count of the experimental account, demonstrating a strong capability for promoting content through social media. It should also be noted that higher engagement points in this study are rounded, as the data scraping process captures only publicly available data

We tried another experiment using the same variables but using entertainment topics as our content focus in our second social media account: Soerabie (**EVA**). In this experiment we produced 10 reel videos content using same variables and the real-time data collected on Sept, 09 2022 using our proposed model and resulted remarkably in an average of 1309.4 virality powers. In this experiment, we also compare the result with 2 accounts (**N1** and **N2**) with similar followers count and 1 account (**HA1**) which has 10 time of the followers count for validation. Our comparison in table 3 shows that our proposed model treated content performs way better compared to **N1**, **N2**, and **HA1**.

Table 3: Second Experiments Result and Comparison

Account Type	EVA	N1	N2	HA1
Followers Count	790	770	801	7928
$e_{1\mu}$	1025650	478	1324	149765
$e_{2\mu}$	8130	73	76	56365
$e_{3\mu}$	236	2	7	26303
$e_{4\mu}$	385	0	3	24539
Power	1309.4	0.7	1.8	32.4

Our third experiments consist of recycling our social media content from the first experiment into another social media platform. We recycle 20 of our reel videos content to Instagram to measure its virality power and ensure the result of the virality. Our experiment shows remarkable results which all of our content performs above 1 point which is the virality power as shown in Table 4.

Table 4: Third Experiments Result

No	Video ID	e_1	e_2	e_3	e_t	Power
1	CisSTjppACd	4500000	551000	663	5051663	7517.356
2	Ciye85tKvgK	1700000	47300	463	1747763	2600.838
3	Cjw11wypi-g	962000	13700	416	976116	1452.554
4	CkRMjhWqZmp	826300	36	5	826341	1229.674
5	CjEbVmYoUIt	504000	14300	198	518498	771.574
6	CjAgSijpt1N	489000	27100	328	516428	768.494
7	CiB1B-EJP-n	448000	8231	301	456532	679.363
8	Ci4M4vIpQcL	316000	5680	231	321911	479.034
9	Ckw1kR0pwiS	310000	9430	128	319558	475.533
10	Cim18pBpyTC	55800	1206	19	57025	84.859
11	Ci1wsZbpPSa	10600	125	3	10728	15.964
12	Ci6gLGDJ4dj	8348	158	6	8512	12.667
13	CjHt_0QJhMO	8348	97	1	8446	12.568
14	Ci-GMmmJYCr	6225	58	4	6287	9.356
15	Ci0H_B7p18O	5554	56	0	5610	8.348
16	CjNsvboJLyh	5183	50	5	5238	7.795
17	Civ1wo5IS5R	3803	33	0	3836	5.708
18	CiketDeppUD	2787	32	3	2822	4.199
19	CiutnY9JoIQ	2290	23	0	2313	3.442
20	CiRaYp2Jjso	2098	24	0	2122	3.158

The highest content power, which is 7517.356, performed better than the first experiment. This could be due to the differences in the social media platforms' algorithms compared to our first experiment. It should also be noted that higher engagement points are rounded in this study because the data scraping process only captures publicly available data. Further in-depth research is needed to examine the differences in virality power across various social media platforms.

Although these experiments focuses on the conceptual and methodological proposal, standard measures such as accuracy, precision, recall, and AUC are appropriate indicators of a good virality prediction. These should be tested in future empirical validations of the framework.

Limitations of the Study

While this study demonstrates the potential of predictive analytics in modeling social media virality, several limitations should be acknowledged. First, the dataset was derived primarily from public data collected through scraping and platform-accessible endpoints. This introduces potential sampling bias, as private interactions and non-public engagement signals were excluded, which may affect representativeness across the wider user base.

Second, the proposed model was validated on selected platforms, particularly TikTok and Instagram. Given that each platform employs distinct recommendation algorithms and user interface dynamics, the findings may not generalize seamlessly to other platforms such as Twitter (X) or YouTube. Future studies should therefore include a broader range of social networks to strengthen external validity.

Third, the model relies heavily on observable engagement metrics as proxies for virality. Although these indicators are widely accepted, they may not capture underlying psychological or cultural drivers of sharing behavior. Integrating qualitative insights from social science, such as motivations for participation and patterns of social influence, would provide a richer explanatory framework.

Fourth, real-time scraping processes are subject to technological and ethical constraints, including changing platform policies and limitations on data access. These factors may affect the continuity of data collection and the replicability of experiments over time.

Finally, although the association rule–based approach improves interpretability, it may oversimplify the complex interactions among features. More advanced machine learning models, such as temporal graph networks or Hawkes process–based predictors, could be incorporated in future research to capture nonlinear dynamics and temporal dependencies.

By acknowledging these limitations, we aim to provide transparency and highlight avenues for future research, ensuring that the findings contribute constructively to both academic inquiry and practical applications.

Conclusion

In conclusion, we have developed a robust model for leveraging social media virality technology to enhance business promotion. This model effectively identifies and utilizes various virality variables, establishes relevant parameters, and has demonstrated remarkable success in increasing engagement.

Our study underscores the significance of social media virality technology in modern marketing strategies. The empirical results affirm the model's effectiveness in generating user engagement and expanding brand awareness. Key to our model's success is the inclusion of real-time content engagement variables, such as retention rates, visual elements, trending topics, hashtags, and audio cues. These factors collectively contribute to notable improvements in content virality and audience engagement. By integrating real-time data collection with advanced analytics and machine learning techniques, our model optimizes social media performance, achieving targeted marketing goals. The continuous evolution and refinement of this model, driven by ongoing data collection and analysis, promises even greater future success in harnessing social media's potential for business promotion.

The robustness of our model was validated through the outstanding virality power observed in our experimental content. Additionally, we developed a generative NLP prototype using the OpenAI GPT API to generate topics, captions, and hashtags, further enhancing our content strategy. The positive outcomes from these experiments, including increased reach, follower growth, and enhanced brand awareness, highlight the efficacy of our approach.

Looking ahead, we remain committed to refining and enhancing our model to fully exploit the capabilities of social media platforms. We are enthusiastic about the future prospects and opportunities that social media presents, and we are dedicated to leveraging its power to drive business success.

Abbreviations

AI: Artificial Intelligence, API: Application Programming Interface, GPT: Generative Pre-trained Transformer, NLP: Natural Language Processing, UKSW: Universitas Kristen Satya Wacana.

Acknowledgement

The authors would like to thank you University of Malaysia Sarawak and Kalimantan Institute of Technology for the funding.

Author Contributions

Riovan Styx Roring: Conceptualization, data collection, data analysis, writing, Chih How Bong: Conceptualization, supervision, suggestions, finalizing paper, Narayanan Kulathuramaiyer: Conceptualization, supervision, suggestions, finalizing paper.

Conflict of Interest

The authors declare no conflict of interest, financial or otherwise.

Declaration of Artificial Intelligence (AI) Assistance

The authors declare that generative artificial intelligence (AI) and AI-assisted technologies were only used to enhance the writing style, grammar, and language clarity of this manuscript. The core intellectual content, including the conception of the study, methodology, analysis, interpretation of results, and conclusions, was developed entirely by the authors. No part of the scientific content, data analysis, or interpretation was generated by AI systems. All responsibility for the accuracy and integrity of the manuscript remains with the authors.

Ethics Approval

Not applicable.

Funding

Funded by Kalimantan Institute of Technology.

References

1. Shiroishi Y, Uchiyama K, Suzuki N. Society 5.0: For Human Security and Well-Being. Computer (Long Beach Calif). 2018;51(7). doi: 10.1109/MC.2018.3011041
2. Fukuda K. Science, technology and innovation ecosystem transformation toward Society 5.0. International Journal of Production Economics. 2020;220:107460. https://doi.org/10.1016/j.ijpe.2019.07.033
3. Asur S, Huberman BA. Predicting the future with social media. Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence. 2010:492–499. https://doi.org/10.1109/WI-IAT.2010.63
4. Moro S, Rita P, Vala B. Predicting social media performance metrics and evaluation of the impact on brand building: A data mining approach. J Bus Res. 2016;69(9). doi: 10.1016/j.jbusres.2016.02.010
5. Van Den Oord A, Kalchbrenner N, Vinyals O, *et al*. Conditional image generation with PixelCNN decoders. In: Advances in Neural Information Processing Systems. 2016. https://proceedings.neurips.cc/paper/2016/hash/b1301141feffabac455e1f90a7de2054-Abstract.html
6. Zhu X, Davidson I. Knowledge discovery and data mining: Challenges and realities. Knowledge Discovery and Data Mining: Challenges and Realities. 2007. https://doi.org/10.4018/978-1-59904-252-7
7. Kaihara T, Kita H, Takahashi S. Innovative Systems Approach for Designing Smarter World. Innovative Systems Approach for Designing Smarter World. 2020. https://doi.org/10.1007/978-981-15-6651-6
8. Lazer D, Pentland A, Adamic L, *et al*. Social science: Computational social science. Science. 2009;323(5915):721–723. https://doi.org/10.1126/science.1167742
9. Lukowicz P, Intille S. Experimental methodology in pervasive computing. IEEE Pervasive Comput. 2011;10(2). https://www.computer.org/csdl/magazine/pc/2011/02/mpc2011020094/13rRUwh80rG
10. Adcock R, Collier D. Measurement validity: A shared standard for qualitative and quantitative research. American Political Science Review. 2001;95(3). https://polisci.berkeley.edu/sites/default/files/people/u3827/APSR2001-Validity.pdf
11. Wang F, Hannafin MJ. Design-based research and technology-enhanced learning environments. Vol. 53, Educational Technology Research and Development. 2005. https://doi.org/10.1007/BF02504682
12. Roring RS, How BC. Towards Society 5.0: A Pilot Study on Costless Smart Transportation Business Model. International Journal of Business and Society. 2022 Mar 31;23(1):73–87.
13. Benda NC, Das LT, Abramson EL, *et al*. "how did you get to this number?" Stakeholder needs for implementing predictive analytics: A pre-implementation qualitative study. Journal of the American Medical Informatics Association. 2020;27(5). doi: 10.1093/jamia/ocaa021
14. DeLone WH, McLean ER. Information Systems Success Measurement. Foundations and Trends® in

Information Systems. 2016;2(1). http://dx.doi.org/10.1561/2900000005
15. Nilashi M, Nilashi M, Samad S, *et al*. Neuromarketing: A Review of Research and Implications for Marketing. Vol. 7, Journal of Soft Computing and Decision Support Systems. 2020. https://jscdss.com/index.php/files/article/view/223/0
16. Morstatter F, Pfeffer J, Liu H, Carley KM. Is the sample good enough? Comparing data from Twitter's Streaming API with Twitter's Firehose. Proceedings of the International AAAI Conference on Web and Social Media. 2013;7(1):400-408. https://doi.org/10.1609/icwsm.v7i1.14401
17. Rizoiu MA, Xie L, Sanner S, Cebrian M, Yu H, Van Hentenryck P. Expecting to be HIP: Hawkes intensity processes for social media popularity. In: Proceedings of the 26th International Conference on World Wide Web (WWW '17). 2017:735-744. https://doi.org/10.1145/3038912.3052650
18. Agrawal R, Srikant R. Fast algorithms for mining association rules in large databases. Proceedings of the 20th International Conference on Very Large Data Bases (VLDB); 1994:487–499. https://dl.acm.org/doi/10.5555/645920.672836
19. Rahm E, Do HH. Data cleaning: Problems and current approaches. IEEE Data Engineering Bulletin. 2000;23(4):3–13.
20. Munzner T. Visualization Analysis and Design. CRC Press; 2014 https://doi.org/10.1145/3721241.3733989
21. Trezza D. To scrape or not to scrape: the post-API scenario and implications on digital research. Frontiers in Sociology. 2023;8:114503. doi:10.3389/fsoc.2023.114503
22. Luscombe A, Dick K, Walby K. Algorithmic thinking in the public interest: Navigating technical, legal and ethical hurdles to web scraping in the social sciences. Quality & Quantity. 2022. https://doi.org/10.1007/s11135-021-01164-0
23. Brady WJ, Jackson JC, Lindström B, *et al*. Algorithm-mediated social learning in online social networks. Trends in Cognitive Sciences. 2023;27(10):947–960. doi:10.1016/j.tics.2023.06.008.
24. Xu Z, Qian M. Predicting popularity of viral content in social media through a temporal-spatial cascade convolutional learning framework. Mathematics. 2023;11(14):3059. doi:10.3390/math11143059.
25. Kozinets RV. Netnography: Doing Ethnographic Research Online. Sage Publications; 2010. https://doi.org/10.2501/S026504871020118X
26. Gorwa R, Binns R, Katzenbach C. Algorithmic content moderation: technical and political challenges in the automation of platform governance. Big Data Soc. 2020;7(1):1–15. doi:10.1177/2053951720919468
27. Bak-Coleman JB, Alfano M, Barfuss W, Bergstrom CT, Centeno MA, Couzin ID, *et al*. Stewardship of global collective behavior. Proc Natl Acad Sci USA. 2021;118(27):e2025764118. doi:10.1073/pnas.2025764118
28. Weng L, Menczer F, Ahn YY. Virality prediction and community structure in social networks. Sci Rep. 2013;3:2522. doi:10.1038/srep02522

How to Cite: Roring RS, Bong CH, Kulathuramaiyer N. From Data Mining to Predictive Analytics: Progress in Understanding and Forecasting Social Media Virality. Int Res J Multidiscip Scope. 2026; 7(1):241-263. DOI: 10.47857/irjms.2026.v07i01.06486