ontributions that explore how statistical methodology, data science, and igence can jointly enhance the analysis and interpretation of business and nic data.

stems from the scientific discussions developed during the international Measuring and Interpreting World Changes with Statistics, Data Science and me from 18 to 20 September 2024. The conference was jointly organised by on for Applied Statistics (ASA), the Department of Statistical Sciences of versity of Rome, and the Italian National Institute of Statistics (ISTAT), with the of several academic and institutional partners. The event provided a ary forum for examining the contribution of statistics, data science, and artificial the understanding of contemporary economic and social transformations.

Issue addresses both theoretical and applied perspectives, focusing on the ability of complex, high-dimensional, and heterogeneous data generated by ons, and digital platforms. Particular attention is devoted to the integration of atistical sources with non-traditional data, including administrative records, and textual information, as well as to the development of transparent and l models capable of supporting evidence-based decision making.

outions collected in this volume investigate advanced methodological empirical applications, and interdisciplinary approaches aimed at improving strategies in areas such as business analysis, labour markets, innovation nd territorial dynamics. By combining classical statistical reasoning with hine-learning techniques, the volume highlights both the opportunities and s associated with the adoption of AI-based tools, including issues related to y, data quality, and responsible use of algorithms.

or researchers, practitioners, and policymakers, this Special Issue provides a coherent overview of current developments at the intersection of statistics, and artificial intelligence, contributing to the ongoing scientific debate on tive methods can effectively support decision-making processes in complex nic contexts.

e is published within the TESI & TEMI editorial series, jointly promoted by ercatorum and the Centro Studi delle Camere di Commercio G. Tagliacarne, heir shared commitment—developed in cooperation with the Association for stics—to fostering high-quality research and scientific dialogue in applied business analytics.

Special issue 1 **Analysing Business Data with Statistics, Data Science, and AI**

TEMI territori economie mercati istituzioni

Dipartimento di Scienze Statistiche
Sapienza Università di Roma

ASA Associazione per la Statistica Applicata

Special issue 1

Analysing Business Data w Statistics, Data Science, a

Editors
Fabio Crescenzi, Luigi Fabbris, Andrea Mazzitelli, Alessandra Righi, Alessandro Rinaldi, Maurizio Vichi

Articles

1986 2026
CENTRO STUDI DELLE CAMERE DI COMMERCIO
Univ Mer

APPROFONDIMENTI SU TWEET IN LINGUA ITALIANA: DISTRIBUZIONI, CONTENUTI, SENTIMENT, MULTIMEDIA E METRICHE DI RETE

INSIGHTS FROM ITALIAN TWEETS: DISTRIBUTIONS, CONTENT, SENTIMENT, MULTIMEDIA, AND NETWORK METRICS

Domenica Fioredistella Iezzi[1], Roberto Monte[2] and Daniele Pasquini[3]

Sommario

Comprendere i meccanismi che guidano la viralità consente di individuare i fattori che plasmano la popolarità e i livelli di coinvolgimento delle principali tendenze nei social su diversi domini. Questo lavoro persegue tre obiettivi principali: (1) analizzare la distribuzione dei messaggi su un ampio campione di tweet; (2) modellare i pattern temporali dei messaggi virali e identificare le principali macro-dimensioni che contribuiscono alla natura virale dei contenuti social; (3) esaminare la struttura di rete degli utenti che hanno pubblicato tali contenuti. La letteratura individua due principali approcci alla previsione delle condivisioni: la predizione della popolarità basata sul contenuto dei messaggi, analizzando i testi e i contenuti multimediali dei post, e quella della popolarità basata sulla struttura della rete. In questo ultimo approccio, si modella la struttura della rete per comprendere come i messaggi vengano condivisi tra gli utenti. La nostra ricerca adotta un approccio a metodi misti, integrando analisi del sentiment, indicatori chiave di performance e metriche di popolarità degli utenti, al fine di caratterizzare le componenti della viralità. Per testare il modello proposto, analizziamo un dataset di 22.155.362 tweet italiani, pubblicati tra il 1° dicembre 2020 e il 12 dicembre 2020.

Abstract

Understanding the mechanisms that drive virality can reveal the factors shaping the popularity and engagement levels of central societal trends and topics. This paper has three main objectives: (1) to analyze message distribution across a large sample of tweets, (2) to model the temporal patterns of viral messages and identify key macro-dimensions contributing to the viral nature of social content, and (3) to exa-

[1] Università di Tor Vergata, Department of Enterprise Engineering "Mario Lucertini", Rome, Italy - e-mail: stella.iezzi@uniroma2.it

[2] Università di Tor Vergata, Department of Civil Engineering and Computer Science Engineering, Rome, Italy - e-mail: roberto.monte@uniroma2.it

[3] Università di Tor Vergata, Department of Enterprise Engineering "Mario Lucertini", Rome, Italy - e-mail: psqdni@hotmail.it

mine the network structure of users who posted this content. The literature identifies two primary approaches to predicting shares: Content-based popularity prediction, which examines textual and multimedia attributes within posts, and Circulation-based popularity prediction, which models the network structure to understand how posts spread among users. Our research employs a mixed-method approach, integrating sentiment analysis, key performance indicators, and user popularity metrics to characterize the components of virality. To test our model, we analyze a dataset of 22,155,362 Italian tweets from December 1, 2020, to December 12, 2020.

Parole chiave: viralità nei social media, big data, rete sociale, analisi del sentiment, modello di regressione.
Keywords: *virality in social media, big data, social network, sentiment analysis, regression model.*

1. Introduction

The number of people using social media to share information is steadily increasing. Social media is a digital platform that connects individuals, enables content creation and sharing, facilitates knowledge exchange, and preserves valuable information for future access (Ghaisani *et al.,* 2017). According to We Are Social and Meltwater (2024), the number of active social media profiles worldwide has surpassed 5 billion, reaching 5.04 billion, over 62% of the global population. This global total has grown by 266 million in the past year, reflecting an annual increase of 5.6%. This remarkable figure shows that, over the last year, the world has averaged an astounding 8.4 new social media users per second. The habits of Italians align with those of other countries around the world. According to the research presented in the report on Italians, a significant amount of time is spent online for various reasons. The primary reasons people access social media are to stay informed about current events and to entertain themselves in their free time (47%), followed closely by the desire to keep in touch with friends and family (45%).

Additionally, there is an increase in the daily time spent on social media and the number of people who report watching video content (91%). This growth is primarily driven by content in the "comedy, memes, and viral videos" category (+3.7%). As a result, certain types of content (posts, tweets, messages, short videos) are particularly engaging and quickly shared with many users.

We talk about virality, which refers to the ability of content to spread rapidly and widely on the internet, often through shares and word of mouth. It is a common phenomenon on social media where a video, post, or meme can go viral, reaching a vast audience quickly. The adjective "viral" comes from the Latin word "poison". According to

Dimmock *et al*. (2016), the discovery of virus's dates to 1892 when Dmitrij Ivanovsky described a non-bacterial pathogen capable of infecting tobacco plants in a paper. Later, the Oxford Dictionary included "viral" among adjectives, defining it as a neologism to indicate something "that spreads particularly quickly and widely, especially through new communication media" or "that tends to spread extensively." In social media, viral diffusion refers to how content quickly spreads across a digital platform. In this case, the "viral" content is spontaneously shared by many people, exponentially increasing its visibility. Messages that are highly retweeted, e.g., are social indicators to evaluate the ability of content (eventually accompanied by video or photos) to spread quickly and widely across social networks and online platforms. This indicator reflects the level of interest, engagement, and social relevance a piece of content generates among users, as sharing is often driven by emotional reactions, timeliness, novelty, or the desire to express one's identity and values. Retweet very popular is, therefore, a social indicator that provides insights into the dynamics of idea dissemination and how specific content reflects and influences collective values, emotions, and interests.

We want to discover the probability distribution of these tweets to understand the characteristics that can make a message go highly retweeted on the platform.

This paper aims to characterize virality by analyzing the distribution of messages from a large sample of tweets. It will model the temporal distribution of viral messages, identify key macro-dimensions contributing to the viral nature of social content, and examine the network structure of individuals who have posted the content. Additionally, the study will investigate the propagation time of a tweet to further understand these dynamics.

The network structure of all social platforms provides information about users and how they are connected through a web of relationships. In this structure, users, with their ties and interactions, such as followers, friends, or direct connections, form a network-like structure (or "grid").

We tested our model on 22,155,362 Italian tweets on various topics collected between December 1 and December 12, 2020. Of these tweets, 6,281,784 received at least one retweet, while 15,873,578 did not. Using a big data dataset, we aimed to validate the model's effectiveness in predicting content popularity across various topics.

The structure of this paper is as follows: Section 2 examines the framework and research directions; Section 3 explores the time for a tweet to become highly shared; Section 4 analyzes sentiment, multimedia elements, and topics within the most retweeted messages in our sample; Section 5 presents our network metrics; Section 6 details regression models for overdispersal count responses and presents our findings; finally, Section 7 concludes with implications and directions for future research.

2. Framework and Research Directions

Various studies have examined the factors influencing the online diffusion of information and electronic word-of-mouth (e-wom) in social networks, highlighting how these processes are analogous to viral contagion mechanisms. Ngo *et al*. (2024) explore the complex relationships among various dimensions of e-wom information, including its credibility, usefulness, adoption, and attitudes toward it. They examine how these factors collectively influence online purchase intentions. Phelps *et al*. (2004) analyze findings from three studies investigating consumer motivations and responses to forwarding emails. The authors discuss the implications for target selection and message creation, providing valuable insights for advertising practitioners looking to implement viral marketing strategies. Additionally, they offer recommendations for future research focused on computer-mediated consumer-to-consumer interactions, highlighting areas of interest for academic researchers.

Several studies (e.g., Berger & Milkman, 2012; De Vries *et al*., 2012; Phelps *et al*., 2004; Kwak *et al*., 2010; Trilling *et al*., 2017) have analyzed the phenomenon of online popularity, which serves as a key indicator of social behavior in digital environments. A variety of metrics can be used to assess popularity, including:

1. Number of Shares: the frequency with which users redistribute the content across their networks.
2. Number of Views: the total number of times the content has been accessed or watched.
3. Engagement: the level of user interaction with the content, encompassing likes, comments, and shares.
4. Growth Rate: the speed at which the content accumulates views and shares over a given time.
5. Reach: the total number of unique users who have encountered the content, either directly or via sharing.

Kim (2018) investigates how social media virality metrics impact perceptions of message influence on oneself and others and intentions to take preventive actions. In this online experiment, participants viewed a Facebook post discussing a health risk, with variations in virality metrics such as the number of likes and shares. The findings reveal distinct effects associated with these metrics: high share counts significantly enhanced perceived message influence on oneself and others and increased intentions to engage in preventive behaviors. Elmas *et al*. (2023) suggested using the ground truth data provided by Twitter's "Viral Tweets" topic to review the current metrics and propose new metrics. Tiago *et al*., 2019 analyze the content promoting tourist destinations. The YouTube platform, which has video content, has proven engaging in this sector.

Bene (2017) addresses the issue of virality in political content messages on Facebook. The results showed that citizens are highly reactive to posts containing negative emotions, text- based posts, personal posts, and those that require action. Virality is mainly facilitated by memes, videos, harmful content, and mobilizing posts, as well as posts containing a request for sharing. Analyzing the most frequently e-mailed New York Times (NYT) articles, Berger and Milkman (2012) found that content virality correlates positively with its positivity and emotional impact, particularly for emotions such as anger, awe, and anxiety, while it is negatively correlated with sadness. Using a sample of German articles, Heimbach and Hinz (2016) replicated their study for the most e-mailed list of Germany's leading news magazine and expanded the analysis to include (1) three additional communication channels and (2) the non-linear relationship between positivity and virality. From a methodological perspective, Avalle *et al*. (2024) provide a large-scale comparative analysis of online conversations across eight social media platforms and over three decades, encompassing more than 500 million comments. Their results demonstrate the persistence of heavy-tailed engagement distributions and invariant toxicity patterns, suggesting that platform design plays a secondary role compared to stable human behavioral dynamics. While we could not replicate BM's findings, our results align with their conclusions across all communication channels. Additionally, we propose that the relationship between positivity and virality exhibits an inverted U-shape, indicating a non-linear pattern.

Our research questions are as follows:

1. How can we characterize virality by examining the message distribution within a large sample of tweets?
2. What are the key macro-dimensions that contribute to the viral nature of social content when studying the temporal distribution of viral messages?
3. How does the network structure of individuals influence the diffusion of highly retweeted content? The probability distribution of retweets offers a comprehensive perspective on message propagation, enabling prediction, optimization, risk management, and informed decision-making.

3. Retweet distribution

To characterize the retweet distribution, we treat the number of retweets received by a single tweet as a count of failures: a tweet with zero retweets is considered to have zero failures, one retweet corresponds to one failure, and so on. This mirrors the conceptual framework of the geometric distribution, which models the number of failures before the first success with probability mass function (PMF) given by:

$$f(n|p) := \mathrm{p}(1-\mathrm{p})^{\mathrm{n}} \quad \mathrm{n} \in \mathbb{N}_0,\ \mathrm{p} \in (0,1),$$

where n is the number of failures before the first success and p is the success probability parameter. On the other hand, the tweets with zero failures vastly outnumber all other tweets. To capture this over-representation of zeros, we adopt a zero-inflated geometric (ZIG) distribution with PMF given by:

$$f(n|\varphi,p) := \begin{cases} \varphi + (1-\varphi)p, & if\ n = 0, \\ (1-\varphi)p(1-p)^n, & if\ n \in \mathbb{N}, \end{cases} \varphi, p \in (0,1),$$

which introduces an inflation parameter φ to increase the probability mass at $n = 0$. Another issue is that each tweet is unique in several respects, and it is unrealistic to assume a constant probability parameter p across all tweets. Instead, we treat p as a random variable, varying across tweets. This approach leads us to model the retweet distribution as a mixture of geometric distributions with different success probabilities. Formally, we assume that our PMF is given by

$$p(n|\vartheta) := \int_0^1 f(n|\varphi,p)\mathrm{g}(p|\vartheta)dp,$$

where $\mathrm{g}(p|\vartheta)$, referred to as mixing distribution, is the density of the success parameter p. Following the structure of "Buy 'Till You Die" (BTYD) models (see Ping *et al*., 2022; Chou *et al*., 2022), we model the mixing distribution of p using a beta density with shape parameters $\alpha, \beta > 0$, that is:

$$\mathrm{g}(p|\vartheta) \equiv \mathrm{g}(p|\alpha,\beta) := \frac{p^{\alpha-1}(1-p)^{\beta-1}}{\mathrm{B}(\alpha,\beta)},$$

where $\mathrm{B}(\alpha,\beta)$ is the beta function. Consequently, a direct computation shows that our candidate distribution for the retweets results in a zero-inflated beta-geometric (ZIBG) distribution given by:

$$f(n|\varphi,\alpha,\beta) := \begin{cases} \varphi + (1-\varphi)\dfrac{\mathrm{B}(\alpha+1,\beta)}{\mathrm{B}(\alpha,\beta)}, & if\ n = 0, \\ (1-\varphi)\dfrac{\mathrm{B}(\alpha+1,\beta+n)}{\mathrm{B}(\alpha,\beta)}, & if\ n \in \mathbb{N}, \end{cases} \varphi \in (0,1), \alpha, \beta > 0,$$

which offers a great flexibility in capturing the heterogeneity of tweet performance. Another advantage of the choice of the beta distribution as the mixing distribution is that it is possible to compute its first three moments in a closed form. This allows us

the application of the method of the moments to determine preliminary estimates of the parameters φ, α, β, through computational procedures for solving nonlinear-equations, and these preliminary estimates can be used as starting points of the computational methods for maximizing the closed form of the log-likelihood function.

It is worth noting that the ZIG distribution could have been replaced by a hurdle geometric distribution (see Cragg, 1971) with PMF given by:

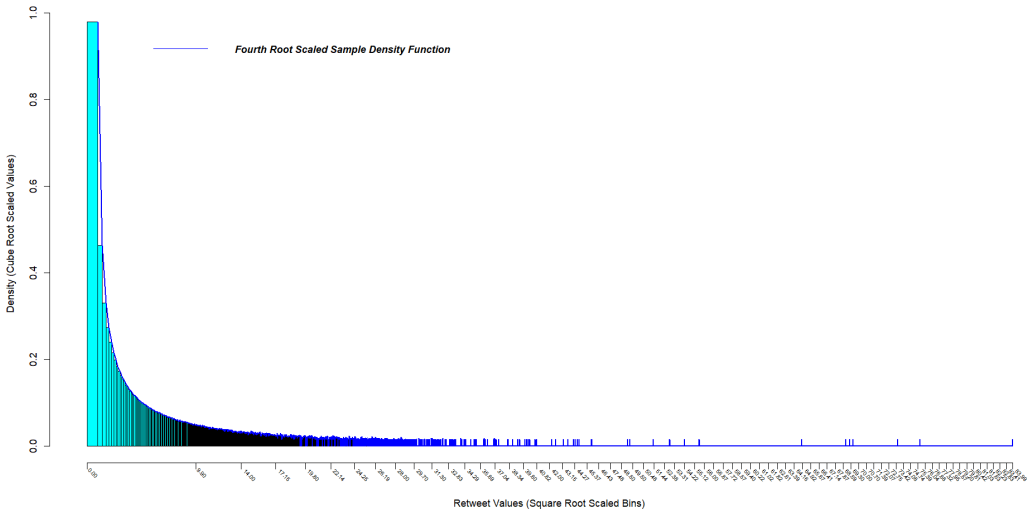
$$f(n|\varphi,p) := \begin{cases} \varphi, & if\ n = 0, \\ (1-\varphi)p(1-p)^{n-1}, & if\ n \in \mathbb{N}, \end{cases} \varphi, p \in (0,1),$$

where φ is the hurdle parameter increasing the probability mass at . In this case, by mixing with the beta distribution we would obtain a hurdle beta geometric (HBG) distribution with PMF given by:

$$f(n|\varphi,\alpha,\beta) := \begin{cases} \varphi, & if\ n = 0, \\ (1-\varphi)\dfrac{\mathrm{B}(\alpha+1,\beta+n-1)}{\mathrm{B}(\alpha,\beta)}, & if\ n \in \mathbb{N}, \end{cases} \varphi \in (0,1), \alpha, \beta > 0,$$

The HBG distribution would offer a simpler structure than the ZIBG distribution, while still providing closed forms for the first three moments. However, for modeling situations with a high number of zero retweets, the ZIBG distribution seems more appropriate to us. This is because the HBG distribution is more suitable for scenarios in which excess zeros stem from a single process, specifically, in our case, some tweets will not be retweeted at all, while all other tweets are guaranteed to be retweeted according to a geometric distribution. In contrast, the ZIBG distribution accommodates scenarios where excess zeros are generated by two processes, in our case, some tweets will never be retweeted, while others may have the potential to be retweeted but do not get retweeted. Considering these key characteristics of the two distributions, we have chosen to use the ZIBG distribution for its superior flexibility.

Figure 1. Density of the Italian retweets from December 1st to December 12, 2020



Fonte: nostre elaborazioni su dati estratti da Twitter

Figure 1 shows the density of retweets in Italian tweets from December 1st to December 12th, 2020.

In addition to modeling the number of retweets, we also consider the virality time, defined as the period during which content spreads rapidly and extensively across a social network or the Internet. This concept encompasses several key aspects: the speed at which the content is shared, the peak moment of its diffusion, and the overall duration of its viral cycle. In our analysis, the unit of measurement for virality time is the hour.

4. Sentiment, multimedia and topics of most rretweeted messages

We analyze the most retweeted tweets, focusing on those with at least 1,000 retweets, representing approximately 1% of all tweets (349,491 tweets). These messages display a strong sentiment bias: 52% carry a negative sentiment, 43% convey positive sentiment, and only 5% are classified as neutral. Additionally, 9% of these tweets include at least one emoji (see Table 1). The sentiment was manually annotated because there were often colloquial expressions or ironic intent, and thus an unsupervised classification based on a dictionary (Liu, 2015) or model-based approach, VADER (Valence Aware Dictionary and sEntiment Reasoner - Hutto and Gilbert, 2014) would not have produced high-quality results. A supervised classification (see Nalini *et al*., 2023) needs a training set, which we can use like Sentiment140 1 to apply machine learning and deep learning algorithms. Still, in those cases, the accuracy of manual classification is superior to that

of automated classification, although it requires a more significant investment of time and human resources. This manual classification has also been used to identify the use of multimedia. We observed that the use of videos and audio is always associated and accounts for 16%, while photos are much more frequent, accompanying 48% of the most retweeted tweets. Table 1 summarizes the sentiment and the use of multimedia tools in the most retweeted tweets.

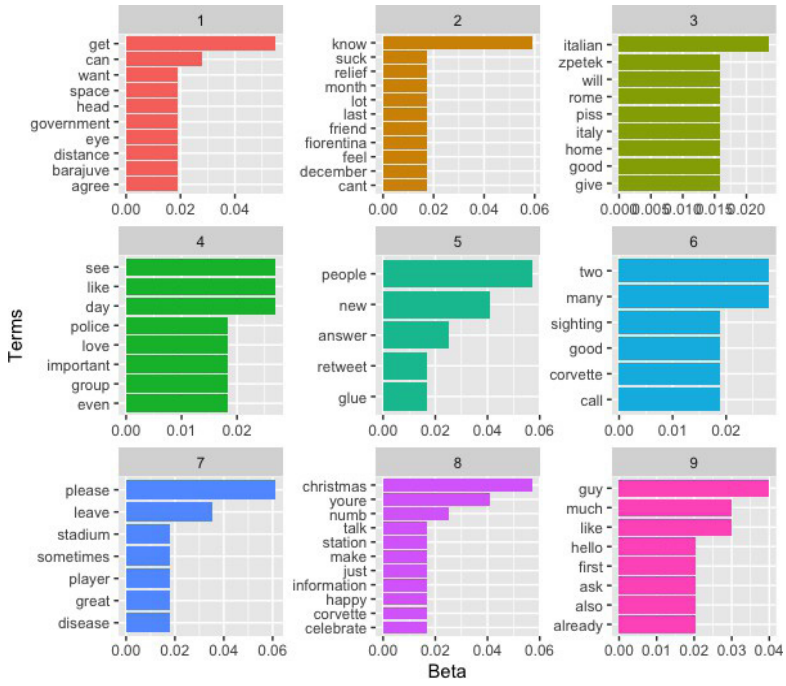
Table 1. Sentiment and multimedial contents in in the most retweeted tweets

Category	Percentage
Positive	43%
Negative	52%
Neutral	5%
Emoje	9%
Photos	48%
Audio + Video	16%

Fonte: nostre elaborazioni su dati estratti da Twitter

To analyze tweet content, we applied the Latent Dirichlet Allocation (LDA) algorithm, a probabilistic model that identifies hidden thematic structures in a text corpus. LDA (see Blei *et al*., 2003) each document (in this case, a tweet) is a mixture of topics, and each topic is a distribution over words. To improve model accuracy, we performed extensive preprocessing, including the removal of URLs, mentions, hashtags, special characters, and numbers, as well as lemmatization. Figure 2 shows the top terms associated with each latent topic extracted through topic modeling (LDA). On the x-axis, Beta represents the estimated probability that a given word belongs to a specific topic. The higher the Beta value, the more representative the word is of that topic.

Figure 2. Top terms for each topic



Fonte: nostre elaborazioni su dati estratti da Twitter

To determine the optimal number of topics, we used perplexity as an evaluation metric: lower perplexity values indicate a better model fit. The final model identified nine topics, each characterized by the most representative terms:

1. Football – tweets centered on football, with frequent terms like "Barajuve", referring to FC Barcelona and Juventus fans.
2. COVID-19 Holiday Restrictions – messages reflecting sadness and isolation due to containment measures during the Christmas period.
3. Lockdown Behavior in Italy – tweets about rediscovered hobbies such as watching Ferzan Özpetek's films during lockdown.
4. Social and Political Engagement – encouragement for solidarity and activism, with expressions like "daleparasempre".
5. Humor and Comedy – lighthearted tweets with jokes and expressions like "hahahaha".
6. Advertising Content – promotional tweets about products, destinations, and cars, such as the Chevrolet Corvette or Braies in South Tyrol.
7. Tributes to Paolo Rossi – condolences and tributes to the legendary footballer.
8. Politics and COVID-19 Policies – criticism of holiday and vaccination policies.
9. Consumer Behavior – tweets about purchases, often online, including platforms like Spotify.

Table 2 reports the main topics identified by the LDA model, together with their thematic interpretation and relative prevalence in the corpus. Most topics account for around 9-11% of the documents, while Topic 9 "Consumer Behavior" is the most represented, covering over one-fifth of the dataset.

Table 2. Topic prevalence and thematic labeling in the Twitter corpus

Topic n°	Description	Dimension (%)
1	*Football*	9.71
2	*COVID-19 Holiday Restrictions*	10.29
3	*Lockdown Behavior in Italy*	8.57
4	*Social and Political Engagement*	9.71
5	*Humor and Comedy*	9.71
6	*Advertising Content*	9.71
7	*Tributes to Paolo Rossi*	10.86
8	*Politics and COVID-19 Policies*	10.29
9	*Consumer Behavior*	21.14

Fonte: nostre elaborazioni su dati estratti da Twitter

Interestingly, the density distribution of followers shows that some highly retweeted tweets originate from accounts with modest followings. However, notable exceptions include Cristiano Ronaldo (110.8M followers), Manchester United (37.7M), and Ibai (13.9M), all linked to the sports sector – highlighting the correlation between virality and sports-related content.

5. Community detection on Twitter/X Heterogeneous Graph

We model user-hashtag interactions using a Heterogeneous Graph $G = \{V, E, \tau, \phi\}$, where V and E are the sets of nodes and edges, respectively, and the functions τ: $V \rightarrow A$ and ϕ: $E \rightarrow R$ map edges in edge types A and nodes in node types R, respectively (retweets RT and hashtag usage M) (Sun *et al*., 2011, 2022). The graph is represented by a symmetric weighted adjacency matrix W with non-negative integer entries, where $W_{i,j} > 0$ if and only if $(i, j) \in E$, and $W_{i,j} = 0$ otherwise. This graph structure captures both relational semantics and content dynamics, enabling a richer representation of social interactions. Retweeting, a key mechanism of information diffusion (Suh *et al*., 2010), reflects the communicative value of content (Cha *et al*., 2010), while hashtag usage supports user visibility (Wang *et al*., 2016) and affiliation with thematic communities (Bruns & Burgess, 2011; Small, 2011; Laucuka, 2018).

To detect communities combining users and hashtags, we tested Louvain (Blondel *et al*., 2008) and Leiden (Traag *et al*., 2019). While both aim to maximize modularity, Leiden was preferred for producing smaller, better-connected clusters and avoiding the oversized partitions typical of Louvain. Other algorithms allowing overlapping communities, such as Conga (Gregory, 2008) and Combo (Sobolevsky *et al*., 2014), were excluded for simplicity.

Using Leiden, we identified over 2 million communities, of which ~90% were singletons. Only 53 had more than 50 nodes. The largest 10 communities (3,795 nodes total) featured distinct thematic areas. For example, communities C_0 and C_9 lacked hashtags (possibly due to the limited time window), while C_1, C_4, and C_6 centered on reality TV (e.g., *Grande Fratello*), and C_3 focused on Turkish TV series. Political content appeared in C_7 and C_8, while COVID-19 and government measures dominated C_2. Community C_5 was more introspective, with users sharing quotes and reflections.

A summary of the top hashtags for the first 10 communities is provided in Table 3.

Table 3. The 10 largest communities by number of nodes, with the most 10 used hashtags by weight

Community	Most 10 used hashtags	# nodes
C_0	N/A	1163
C_1	gfvip, verissimo, secondavita, oppininstudio	676
C_2	mes, conte, covid19, natale, dpcm, m5s, salvini, governo, vaccine	539
C_3	canyaman, özgegürel, mrwrong, produawards2020mrwrong, produawards2020canyaman, canyamanmanoftheyear2020, produawards2020, produawards2020özgegürel, wemissyouözgegürel, bayyanlışrewind	267
C_4	rosmello, dayane, rosalinda, rosmelloilnostrogf, rosmellosempreconvoi	238
C_5	ventaglidiparole, buongiorno, avreivoluto, untemaalgiorno, unsogno, buongiornoatutti, uninvitoa, tuttequellecoseche, cosaèsuccesso, buonanotte	212
C_6	gregorelli, zorzelli, zorpini, pierpaolopretelli, elisabettagregoraci, gregorando, gregorellidellanotte	189
C_7	renzi, report, meloni, lega, reportrai3, fontana, berlusconi, fdi, lanotizia	174
C_8	ottoemezzo, gruber, italiaviva, boschi, philipmorris, casaleggio, cinquestellopoli	169
C_9	N/A	168

Fonte: nostre elaborazioni su dati estratti da Twitter

6. Regression models for over-dispersed count response

To investigate the factors influencing retweets, we applied various regression models (see Cameron and Trivedi, 1990; Korosteleva, 2018, using explanatory variables such as sentiment, the use of multimedia elements like photos, videos, and audio, the number of followers, likes, and network statistics. Retweet data, although they are count data, exhibit overdispersion, meaning that the variance exceeds the mean. For this reason, standard Poisson regression models may not be appropriate. This is because Poisson regression assumes that the mean and variance of the response variable are equal, which is not the case in situations of overdispersion. In this case, we can also use the Zero-Truncated Negative Binomial (ZTNB) regression model, which is used when the response variable is count data that is strictly positive (i.e., there are no zero counts). This model is particularly useful in scenarios where the occurrence of zeros is impossible or does not make sense, such as the number of times an event occurs after it has been triggered. Response Variable: The response variable Y follows a zero-truncated negative binomial distribution. We apply three regression models: Poisson (POIS), Negative Binomial (NB), and ZTNB regression model, using eight distinct models for each type by adding one regressor at a time. Specifically, the regressors encompass two categories of information: content (including audio/video presence, photos, emojis, sentiment, and topics) and network characteristics of the tweeters (followers, hashtags, and clusters).

Table 4. Results Regression models: POIS, NEGB, ZNEGB

Akaike Information (AIC)	POIS	NB	ZTNB
audio/video	137769	1677	1814
audio/video + photo	135383	1676	1813
(audio/video + photo + emoje) = multimedia	121399	1653	1793
multimedia+sentiment	112474	1652	1788
multimedia+sentiment+topic	59769	1633	1790
multimedia+sentiment+topic+follower	59014	1634	1635
multimedia+sentiment+topic+follower+hashtag	59014	**1630**	1631
multimedia+sentiment+topic+follower+hashtag+cluster	59016	1632	1633

Fonte: nostre elaborazioni su dati estratti da Twitter

To identify the most effective statistical models suited to our data, we utilized the Akaike Information Criterion (AIC) for comparison. As demonstrated in Table 4, the NB regression model outperforms all other models when the explanatory variables in-

clude multimedia, sentiment, topic, followers, and hashtags. Additionally, the ZTNB model shows improvement and approaches NB values when network measures are incorporated. These results underscore the benefits of a mixed approach, suggesting that incorporating both content and network variables enhances model performance.

7. Conclusions

This study offers a comprehensive examination of virality on Twitter, showing that the diffusion of messages is shaped by both content features and network structures. The retweet distribution follows a zero-inflated beta-geometric model, which captures the heterogeneity of tweet performance. Measuring virality time in hours provides a more accurate description of the life cycle of viral messages, reflecting the rhythm of their acceleration, peak, and decline.

Content analysis reveals a slight dominance of negative sentiment among the most retweeted tweets, while positive messages remain highly represented and neutral tones are marginal. Multimedia plays a significant role, with photos present in nearly half of the most viral tweets and videos or audio reinforcing engagement despite their lower frequency. Topic modeling highlights a heterogeneous set of themes, with consumer behavior, sports, politics, and COVID-19 emerging as the most influential areas of discussion. Notably, virality does not depend exclusively on large audiences: while sports celebrities and organizations dominate the extremes, smaller accounts can also generate substantial diffusion when content resonates strongly.

Community detection shows that online conversations are organized into a limited number of large thematic clusters, primarily around politics, entertainment, and the pandemic. Finally, regression analyzes confirm that standard Poisson models are inadequate, while Negative Binomial and Zero-Truncated Negative Binomial models perform better, especially when combining content and network predictors. This mixed approach consistently improves explanatory power and predictive accuracy.

In conclusion, virality emerges as the outcome of complex interactions between message design, network configuration, and temporal dynamics. Our findings underline the need for integrated models that account for these dimensions simultaneously, providing insights for both theory and practice. Future research should extend the analysis to multiple platforms, investigate the amplifying role of algorithms, and refine sentiment detection methods to better capture nuances such as irony and sarcasm.

Bibliografia

Avalle, M., Di Marco, N., Etta, G., Sangiorgio, E., Alipour, S., Bonetti, A., Alvisi, L., Scala, A., Baronchelli, A., Cinelli, M., & Quattrociocchi, W. (2024). Persistent interaction patterns across social media platforms and over time. *Nature, 628*(8008), 582-589. https://doi.org/10.1038/s41586-024-07229-y.

Bene, M. (2017). Go viral on the facebook! Interactions between candidates and followers on Facebook during the Hungarian general election campaign of 2014. *Information, Communication & Society, 20*(4), 513-529.

Berger, J., & Milkman, K. L. (2012). What makes online content viral? *Journal of Marketing Research, 49*(2), 192-205.

Blei, D. M., & Lafferty, J. D. (2007). A correlated topic model of science. *The Annals of Applied Statistics, 1*(1), 17-35. https://doi.org/10.1214/07-AOAS114.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3(Jan), 993-1022.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment, 2008*(10), P10008.

Bruns, A., & Burgess, J. (2011). The use of Twitter hashtags in the formation of ad hoc publics. *Proceedings of the 6th European Consortium for Political Research (ECPR) General Conference 2011*, 1-9.

Cameron, A. C., & Trivedi, P. K. (1990). Regression-based tests for overdispersion in the Poisson model. *Journal of Econometrics, 46*(3), 347-364.

Cha, M., Haddadi, H., Benevenuto, F., & Gummadi, K. (2010). Measuring user influence in Twitter: The million follower fallacy. *Proceedings of the International AAAI Conference on Web and Social Media, 4*(1), 10-17.

Chou, P., Chuang, H. H.-C., Chou, Y.-C., & Liang, T.-P. (2022). Predictive analytics for customer repurchase: Interdisciplinary integration of buy till you die modeling and machine learning. *European Journal of Operational Research, 296*(2), 635-651.

Cragg, J.G. (1971). Some Statistical Models for Limited Dependent Variables with Application to the Demand for Durable Goods. *Econometrica, 39(5), 829-844,* https://doi.org/10.2307/1909582.

Dimmock, N., Easton, A., & Leppard, K. (2016, January). *Introduction to modern virology* (7th ed.). Blackwell Publishing.

Elmas, T., Stephane, S., & Houssiaux, C. (2023). Measuring and detecting virality on social media: The case of Twitter's viral tweets topic. *Companion Proceedings of the ACM Web Conference 2023*. https://doi.org/10.1145/3543873.3587373.

Ghaisani, A. P., Handayani, P. W., & Munajat, Q. (2017). Users' motivation in sharing information on social media. *Procedia Computer Science, 124*, 530-535. https://doi.org/10.1016/j.procs.2017.12.186.

Gregory, S. (2008). A fast algorithm to find overlapping communities in networks. *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 408-423.

Heimbach, I., & Hinz, O. (2016). The impact of content sentiment and emotionality on content virality. *International Journal of Research in Marketing, 33*(3), 695-701.

Hutto, C., & Gilbert, E. (2014). VADER: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the International AAAI Conference on Web and Social Media*, 94-100.

Kim, J. W. (2018). They liked and shared: Effects of social media virality metrics on perceptions of message influence and behavioral intentions. *Computers in Human Behavior, 84*, 153-161.

Korosteleva, O. (2018, December). *Advanced regression models with SAS and R*. CRC Press - Taylor Francis Group. https://doi.org/10.1201/9781315169828.

Kulkarni, S., & Rodd, S. F. (2020). Context aware recommendation systems: A review of the state of the art techniques. *Computer Science Review, 37*, 100255. https://doi.org/10.1016/j.cosrev.2020.100255.

Laucuka, A. (2018). Communicative functions of hashtags. *Economics and Culture, 15*(1), 56-62.

Liu, B. (2015). *Sentiment analysis: Mining opinions, sentiments, and emotions*. Cambridge University Press. https://doi.org/10.1017/CBO9781139084789.

Nalini, C., Dharani, B., Baskar, T., & Shanthakumari, R. (2023). Review on sentiment analysis using supervised machine learning techniques. In A. Abraham, S. Pllana, G. Casalino, K. Ma, & A. Bajaj (Eds.), *Intelligent systems design and applications* (pp. 166-177). Springer Nature Switzerland.

Ngo, T. T. A., Bui, C. T., Chau, H. K. L., & Tran, N. P. N. (2024). Electronic word-of-mouth (eWOM) on social networking sites (SNS): Roles of information credibility in shaping online purchase intention. *Heliyon, 10*(11).

Phelps, J. E., Lewis, R., Mobilio, L., Perry, D., & Raman, N. (2004). Viral marketing or electronic word-of-mouth advertising: Examining consumer responses and motivations to pass along email. *Journal of Advertising Research, 44*(4), 333-348.

Small, T. A. (2011). What the hashtag? A content analysis of Canadian politics on Twitter. *Information, Communication & Society, 14*(6), 872-895.

Sobolevsky, S., Campari, R., Belyi, A., & Ratti, C. (2014). General optimization technique for high-quality community detection in complex networks. *Physical Review E, 90*(1), 012811.

Suh, B., Hong, L., Pirolli, P., & Chi, E. H. (2010). Want to be retweeted? Large scale analytics on factors impacting retweet in Twitter network. *2010 IEEE Second International Conference on Social Computing*, 177-184.

Sun, Y., Han, J., Yan, X., Yu, P. S., & Wu, T. (2011). PathSim: Meta path-based top-k similarity search in heterogeneous information networks. *Proceedings of the VLDB Endowment, 4*(11), 992-1003.

Sun, Y., Han, J., Yan, X., Yu, P. S., & Wu, T. (2022). Heterogeneous information networks: The past, the present, and the future. *Proceedings of the VLDB Endowment, 15*(12).

Tiago, F., Moreira, F., & Borges-Tiago, T. (2019). YouTube videos: A destination marketing outlook. *Strategic Innovative Marketing and Tourism: 7th ICSIMAT, Athenian Riviera, Greece, 2018*, 877-884.

Traag, V. A., Waltman, L., & Van Eck, N. J. (2019). From Louvain to Leiden: Guaranteeing well-connected communities. *Scientific Reports, 9*(1), 1-12.

Wang, R., Liu, W., & Gao, S. (2016). Hashtags and information virality in networked social movement: Examining hashtag co-occurrence patterns. *Online Information Review, 40*(7), 850-866.

We Are Social & Meltwater. (2024). *Digital 2023 global overview report*. Retrieved from https://datareportal.com/reports/digital-2024-global-overview-report.