*Journal of Statistics & Management Systems*
ISSN 0972-0510 (Print), ISSN 2169-0014 (Online)
Vol. 28 (2025), No. 4, pp. 761–793
DOI : 10.47974/JSMS-1379

# Predicting the success of online news about movies with Google Analytics and Twitter : A machine learning methodology

Víctor Yeste *
*Department of Applied Statistics and Operational Research and Quality*
*Universitat Politècnica de València*
*Valencia 46022*
*Spain*

*and*

*School of Science, Engineering and Design*
*Universidad Europea de Valencia*
*Valencia 46010*
*Spain*

Ángeles Calduch-Losa †
*Department of Applied Statistics and Operational Research and Quality*
*Universitat Politècnica de València*
*Valencia 46022*
*Spain*

José Antonio Ontalba-Ruipérez §
Jorge Serrano-Cobos ‡
*Department of Audiovisual Communication, Documentation and History of Art*
*Universitat Politècnica de València*
*Valencia 46022*
*Spain*

---

* *ORCID-ID:* **0000-0002-3660-8347**
† *ORCID-ID:* **0000-0001-8287-7823**
§ *ORCID-ID:* **0000-0002-2071-6108**
‡ *ORCID-ID:* **0000-0002-4394-4883**

---

* *E-mail:* **vicyesmo@upv.es (Corresponding Author)**
† *E-mail:* **mcalduch@eio.upv.es**
§ *E-mail:* **joonrui@upv.es**
‡ *E-mail:* **jorserc2@har.upv.es**

© Taru Publications

**Abstract**

This paper presents a machine learning methodology for predicting the success of online movie news with web analytics data from Google Analytics and social media analytics data from Twitter (now known as X). This methodology is built in two phases. The first consists of segmenting the data with the different categories that apply to the articles (in this case, all news published, news about movies, and news with a trailer) and performing a multiple linear regression to extract a prediction equation for each success variable. The second phase includes validating the prediction equations with test data, which helps to select the most reliable prediction depending on its accuracy. This methodology has shown that it can account for some of the variability of the success variables and can make their prediction. It provides a basis for future research to improve accuracy and test new variables that can enrich this analysis. This article also helps editorial teams make better data-driven decisions, such as more efficient resource planning or optimising articles to achieve a specific goal.

***Subject Classification:*** *62J05, 62P25.*

***Keywords:*** *Online success, Online news, Digital journalism, Web analytics, Google Analytics, Social media, Social media analytics, Twitter, X, Success prediction, Multiple linear regression, MLR, Statistics, Machine learning.*

## 1. Introduction

This study is framed within the field of cyber journalism or digital journalism, which, according to Díaz Noci [14], consists of the publication of current information through online media using journalistic texts that take advantage of the hypertextuality, multimedia, interactivity, frequency of updating and form of online content [14]. In addition, online journalism has greater immediacy compared to print media, as it often creates a more vivid sense of being present at the news event locations for that consuming news digitally rather than in print format [39].

These characteristics of cyberjournalism make it possible to monitor consumer behaviour/consumption, which gives rise to the possibility of analysing and predicting the probability of success of the content that could be published, with success being understood as the fulfilment or improvement of metrics related to the objectives set by the authors or the

**Summary Statement of Contribution:** This paper has not only demonstrated that some of the information in the success indicators can be predicted but also encourages more granular and specific analyses. It has approached the success of digital content from a new perspective, as it has opted to design a methodology that can be integrated into the decision-making process of the editorial team of a medium and that can also be adapted to other indicators and use cases.

editorial team of the digital media in which they are published. These metrics can be economical, strategical, and even communicational, as aspirational talk and user engagement are essential in corporate social responsibility (CSR), according to Araujo and Kollat [4]. This engagement can be elusive, so some papers approach this by researching Twitter parameters and analysing trends [36], or looking for network map patterns in the context of news diffusion [1], or predicting user behaviour to identify marketing strategies by using machine learning algorithms [8].

This analysis would help in decision-making in the content creation strategy, optimising the work of the journalist and the rest of the editorial team. A greater predictive capacity of the success of the published content allows for increasing the audience and the profitability of information consumption, as Schudson [47] points out, while helping to know the most effective practices in increasingly varied journalism.

This article focuses on the performance of the content on the internet, specialising in journalistic and breaking news content. It has been found that many researchers have approached the problem in search of the most appropriate indicators but always confined to one area and even to a few specific indicators, without adapting to the analysis of content in various areas and to the content strategy of a medium, with its needs and peculiarities. For example, some studies have found that the trustworthiness, expertise, and attractiveness of a source and the volume of multimedia elements all play a significant role in influencing the likelihood of information being retweeted in the context of emergency events and their diffusion [32]. There is much academic work on trend analysis and event detection, even to predict them. When studying success, the analysis of the entire website, the social network account, or a single post on the social network is usually carried out. However, research has yet to be discovered on measuring from the point of view of a specific content, let alone predicting its analytical parameters.

This makes it necessary to develop a methodology to optimise the strategy of content publication and sharing on social networks. This methodology would be cybermetric, consisting of the measurement, study, and quantitative analysis of online information using bibliometric, scientometric, and information techniques [49]. The advantages of incorporating a methodology of these characteristics in the editorial process are numerous, such as the optimisation of resources to which to allocate the preparation of an article, the choice or discarding of content, the optimisation of content to meet the trend to which it could form part and even the optimisation of digital communication on social networks, as

in the case of Twitter (now known as X). This part of the process is framed in Graefe's [19] news generation model in the step of prioritising information, as can be seen in the Figure 1:

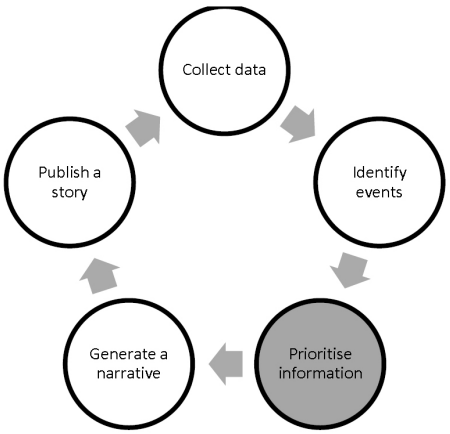



**Figure 1**

**News generation model. This study is framed by information prioritization [19]**

This study has approached the research problem of predicting success of digital content from a new perspective. The goal is to develop a methodology seamlessly integrated into the editorial team's decision-making workflow and easily adapted to different indicators and contexts, ensuring that resources are invested more strategically.

## Objectives

This approach addresses a cybermetric methodology to predict an online publication's success based on data obtained from already published content and trends arising from the public's information needs, which assigns the corresponding priority to the resources of the online medium. The specific objectives are, therefore, the steps necessary to design a methodology and validate it:

(1) Identify the indicators that allow the development of a predictive analysis of online content.

(2) Select an example that can be extrapolated to the population under study.

(3) Describe the extraction of data and its processing to obtain the indicators.
(4) Determine what statistical procedures, tools, and reports are necessary.
(5) Use regression analyses to develop predictive models for the chosen measures of success.
(6) Evaluate the predictive models with test data to determine their accuracy and effectiveness.

## Related Study

*Dissemination of journalism on Twitter*

Within disseminating journalistic content on the web, Twitter is seen as one of the primary news sources in real-time. According to Hermida [21], the number of US users who turn to this social network for breaking news was already twice as high as on Facebook in 2016. This trend could also be seen in a Pew Research Center survey, in which 68% of Americans admitted using Twitter or Facebook to read the news [25]. Peñafiel Sáiz [42] asserts that Twitter is, therefore, the social network that has had the most significant impact on Journalism 2.0, as it involves the audience but also the measurement of the success of news content. This is why, for many Twitter users, the main functionality of the social network is to consume news, as Fox and Lenhart [16] found that users have a particular predilection for reading online newspapers as opposed to printed newspapers (76% compared to 60% of non-Twitter users) and using their mobile phones to do so (8% compared to 1% of non-Twitter users). They also read much more blog content (57% have ever read a blog compared to 29% of non-Twitter users). A decade after the creation of Twitter, roughly three-quarters of journalists in Western countries maintain an active presence on Twitter [30], and, according to Orellana-Rodriguez and Keane [40], nearly all media outlets utilise Twitter as a platform to share their content and engage directly with their audiences, so this social network has become an essential tool for journalists.

*Measuring the success of web content*

Considering the analysis of journalistic content, web analytics is valuable, according to the decision-makers interviewed in a study by Belair-Gagnon and Holton [7], from more of an economic viability than a

news production value point of view. That is, it helps improve the visibility of good content and can help form a sustainable business model. This results in the creation of a new form of journalism, which Carlson [11] calls 'measurable journalism', which includes a cultural and material shift in digital platforms, as they can measure real-time, individualised, and quantitative audience data and consumption.

Web usage mining involves employing data mining methodologies to detect behavioural patterns within web data, ultimately enhancing the comprehension of user needs and guiding improvements to web applications. Zumstein and Kaufmann [62] further add that web analytics is used not only for web optimisation but can also improve web quality, develop digital marketing, optimise search engine optimisation (SEO), the profitability of digital businesses, etc.

*Measuring success on Twitter*

When analysing Twitter activity, it is important to study why and how users use Twitter. According to Java et al. [24], the users' intention of use can be categorised into daily conversation, conversations, information exchange, and community structure. Under this intention, users take on the roles of information source, friend, or information seeker, being considered active if they have published at least one tweet over a week. On the other hand, they would be regarded as retained if they retweet at least once in the following weeks. Java et al. [24] found that approximately half of the users were active, and half had retweeted the following week.

In an article by Teevan et al. [52] comparing web search with Twitter search, it was found that users' intention when using the search tool within Twitter is primarily to obtain timely and current information, including breaking news, live updates, and trending topics, as well as details associated with specific individuals—such as personalised content for the searcher, data from other users of interest, or opinions in general. Twitter search queries are shorter, more influenced by what is currently trending, and less likely to shift during the user's browsing session. According to the same authors, this is partly because Twitter users' intention when searching is to review results related to the query. In contrast, they usually search for more information about the query on a website. All this is related to the nature of content on Twitter, which is more conversational and in real-time than content from a website, where facts and navigational content are more common. Therefore, for Kaushik [26], one of the keys to social network analytics is to measure whether one

is participating in the channels optimally. That is, if it can capture attention, make people enjoy it, provoke the desire to share, start conversations, carry out specific actions, etc.

Rodríguez Martínez [44] provides the following five indicators for measuring social networks: visibility or reach thanks to the traffic received by the website; the number of visits, followers, subscribers, clicks, incoming links; interaction through tweets, retweets and 'likes' received by users; influence according to who has shared the content and how much, number of retweets and number of lists; loyalty and engagement or degree of commitment thanks to the quality of the tweets received regardless of whether they are positive or negative, as well as the links shared, all through manual analysis; and popularity measured by the number of followers. Another approach by other researchers has been analysing the tweets' content using sentiment analysis, showing that the content of the tweets matters even in objectives outside the social media analysed, like TV consumption [48]. Monitoring social media can also be helpful at identifying threat content shared, a need that can exists in many countries depending on the existing political and economic instability [2].

*The dissemination of news on Twitter*

From the outset, Twitter has placed great importance on trends, which initially originate from emerging terms and increase in the frequency of use over time, as they adapt their communication behaviour to the interests of their audience. However, the informational value and other inherent criteria of the messages remain important to them when sharing information [28], like newsworthiness and credibility of the information conveyed [12]. According to [29], most of the trends on Twitter (approximately 85%) were breaking news or news with a longer run, published by traditional media sources, and retweeted on Twitter, generating the trends [5]. This has led some users to create fake information articles or fake news that spread fast, induce false trends, and can potentially create damage to society. Some researchers have detected these misleading articles automatically with machine learning techniques, such as SVM, Logistic Regression, and Naive Bayes [35].

Zubiaga et al. [61] tackled the classification of trends using typologies and without the need for external data. The typologies proposed are news, ongoing events, memes, and commemoration events. To do so, they used the patterns associated with fifteen social characteristics of each type of trend: depth of retweets, retweet ratio, number of hashtags, length of

tweets, exclamation marks, questions, links, repetition of topics, replies, speed of propagation, diversity of users who authored, variety of users who retweeted, diversity of hashtags, diversity of language and diversity of vocabulary.

A different approach is message popularity prediction. For example, Hong et al. [22] proposed a method that measured the number of future retweets to determine what factors influence information propagation on Twitter. They first developed a binary classification model, trained on positively and negatively received posts, to gauge their retweet potential. Next, they applied a multi-class classification model to estimate the possible range of retweet volumes for the new content. On the other hand, Huang et al. [23] tried to predict the active period of trends, considering the active period of popularity evolution as the length of time the content receives continuous attention from users.

## Materials and Methods

This paper presents applied research in a particular and well-defined field, validating it with a use case in which news about cinema is published. The research approach has been quantitative, using various procedures based on measuring the success of publications on Twitter, analysing trends, and web analytics about the medium studied. It has used experimental and longitudinal research using a digital medium in a specific time interval.

The methodology used in this work consisted of the steps shown in the flowchart of Figure 2:

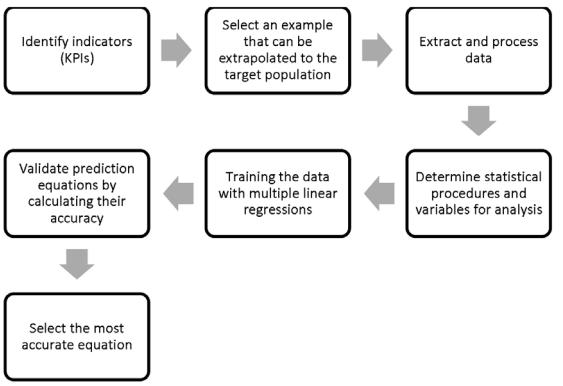



**Figure 2**

**Process flow diagram of the methodology**

**Identify the indicators that allow the development of predictive analysis of online content**

Due to the need to integrate data from two separate areas, such as web analytics and the analysis of shares and related topics on Twitter, we have opted for customised programming that accesses both the Google Analytics v4 reporting API and the Twitter Standard API, always within the limits of these.

Once this data is obtained, and to avoid making too generic measurements, it is necessary to select key performance indicators (KPIs) metrics to quantify objectives and help define and measure progress towards what the business needs.

*Web analytics indicators*

The methodology followed for web analytics from a theoretical point of view is based on five phases, explained in Intriago, Fernández, and Gamboa [3] the definition of objectives and KPIs, the implementation of the appropriate tools and measurement; reporting thanks to a customised user interface; analysis based on the initial objectives; and optimisation through tasks or strategies to improve the website [3].

According to Noguera [38], audience loyalty is measured by page views and time spent on the site. Other media, such as e-intelligent.es (2014), have suggested the following metrics: to measure content consumption, the number of pages viewed, average time on page, bounce rate, number of pages viewed per visit, and number of unique users would be helpful; to measure reach and dissemination, social media shares, number of links received and number of conversations generated by the medium or article would be useful; conversions would be measured by the number of conversions caused by the content strategy, newsletter subscriptions, downloads of essential files and sales generated if the website has them [15].

Google Analytics Reporting API v4 exports data in JSON format, which can then be processed to collect and sort information from its main components: user data, session data, page view data, and Adsense data, among others. User data represents information about each user, with a unique Client ID (referring to the browser itself), a device, a browser, and a user type (new or returning). Session data provides duration, number of searches, page views, etc. Pageview data is data on browsing information for each session, such as time spent on each page browsed when a page was accessed [13].

From this tool, it has been possible to extract the analytical information referring to the content on the website, using the URL of the page as a dimension: ga:pagePath, as shown in Table 1.

**Table 1**
**Web analytics indicators**

| Type | Indicator |
|---|---|
| Acquisition | Unique pageviews (total) |
| | Unique pageviews (average) |
| | Entries (average) |
| Behaviour | Duration on page (average) |
| | Exit rate (average) |
| | Pages per session (average) |
| Results | Ad impressions per session (total) |
| | Ad impressions per session (average) |
| | Percentage of viewable ad impressions (average) |
| | Ad click-through rate (CTR) (average) |
| | eCPM: Estimated Adsense revenue / 1000 pageviews (average) |

*Analytics indicators of the Twitter account of the analysed medium.*

Sharing personal and daily information in digital journalism has made it feasible to automatically extract Twitter users' chronological timelines from their tweets collection [31]. Since, as Ryan, Hazlewood, and Makice [45] found, 85-90% of users leave their Twitter profile public, other services can access and process this data thanks to Twitter's open API [45].

The Twitter Standard API was used to extract the information related to the analytics of both the content of the media outlet's Twitter account and the analysis of Twitter trends. This is the standard and free version of the official Twitter API with the possibility of making basic queries of the last seven days, filtering, obtaining a sample, accessing the public information of user accounts, obtaining trends from a specific location, etc. [56]. Except for the Streaming API, the rest of the Twitter API functionalities act according to REST (Representational State Transfer) design principles, returning responses in JSON format via HTTP [59]. There is a maximum number of requests in 15-minute windows, so Twitter

recommends using caching, prioritization of active users, and filtering or less frequent search queries with no results [57].

When analysing the tweets published by the media account, use was made of the 'statuses/user_timeline' function, which allows up to 900 requests every 15 minutes as a user, 1,500 every 15 minutes as an application, and 100,000 daily requests. It returns a collection of up to 3,200 of the most recent tweets from a particular user, including native retweets from other users [55]. These maximums affect the requests made in the data collection, as temporal maximums have slowed down the procedure with queries that respect the windows allowed by the API, and others, such as the maximum of most recent tweets, have limited the analysis to that collection of tweets.

The analysis of tweets published by the media account has been carried out with the indicators found in Table 2.

**Table 2**

**Analytics indicators in the Twitter account of the analysed media.**

| Type | Indicator |
|---|---|
| Amplification | Retweets (total)<br>Retweets per tweet (average) |
| Praise | Favourites (total)<br>Favourites per tweet (average) |

*Trend analysis indicators of the topics related to the articles.*

The utilities of Twitter data mining are very numerous, from monitoring political activity, detecting earthquakes and flu outbreaks, and detecting false alarms by not raising tweets about them to sociological issues of many kinds since analysing tweets is much less costly in time and money than conducting surveys, according to Savage [46]. This information dissemination, according to Yang and Counts [60], has the following main properties:

- **Velocity:** how followers are influenced and interact with the information, whether by retweeting, replying, or mentioning the initial tweet with their tweets. It, therefore, boils down to analysing if and when someone reacts to the tweet somehow.
- **Scale:** how many users who are followers of the source of information talked about the same topic, i.e., users linked in the first degree of following.

- **Rank:** how much influence the tweet has had according to the distance to the receiver in the dissemination chain, following the degrees of followers until reaching the highest one [60].

To collect data related to the trends of the topics, the Streaming API has been used, the one referring to the tracking of tweets in real-time, consisting of an access that allows up to 400 keywords and up to 5,000 users to be followed. The standard search API allows simple queries of the indexes of the most recent or popular tweets, limited to the last seven days. This functionality supports 180 requests in 15-minute windows if done from a user's point of view and 450 requests in 15-minute windows in the case of a platform-approved application [58]. The 'followers/ids' function has also been used, which returns the set of followers of a paid user in groups of 5,000 with a limit of 15 requests every 15 minutes [54]. These API limitations have affected the methodology in the same way as in the previous section. The indicators used to analyse trends of the topics are shown in Table 3.

**Table 3**

**Trend analysis indicators**

| Type | Indicator |
|---|---|
| Amplification | Tweets (total) |
| | Retweets (total) |
| | Retweets per tweet (average) |
| | Percentage of followers who have posted tweets on the same topic |
| Praise | Favourites (total) |
| | Favourites per tweet (average) |
| Authority | Followers of tweet author (average) |
| | Tweets originating from the tweet's author's account (average) |
| | Age of the tweet's author's account (average) |
| Content | URL inclusion in the tweet (0 / 1) |

**Select an example that can be extrapolated to the study population**

The analysis of results has been carried out by taking as the universe the news items published in a digital medium. The digital medium hellofriki.com has been accepted as a use case due to the availability of

access to all the information necessary to carry out this study. It is a website with informative online entertainment content, with editorial lines on cinema, series, video games, literature, and comics. It has been active since 2009 and has published over ten thousand articles. In 2020, the year in which the data for this study was taken, it had annual traffic of more than 210,000 users and almost 315,000 page views, so it serves as a sample of the population of websites that publish cinema news. This work can also be seen as a testing ground for similar areas.

As an object of study, we have considered the cinema news published during three calendar months in this medium, between 23 September and 22 December 2020. All articles published in that period and trailer news were also considered, as they are thematically related to cinema and are the most topical subpopulations with enough articles to draw statistical conclusions, as the minimum number of articles to be able to draw valid conclusions is 30 [41].

*Describe the data extraction and processing to obtain the indicators.*

The data collection process for each article takes place at two points in time, as can be seen at Figure 3:

- On the day of the article's publication, trend analysis data about the previous seven days is collected.
- Fourteen days later, the data from the content analytics on the website, on the Twitter account of the media outlet, and the trend analysis data are collected, the latter belonging to the seven days before this moment. This period has been chosen because it has been considered sufficient to take robust and valuable data for prediction.

In this research, a theme has been considered a keyword or set of keywords thematically related to a published article. For this purpose, the tags inserted by the articles' authors have been used to analyse trends

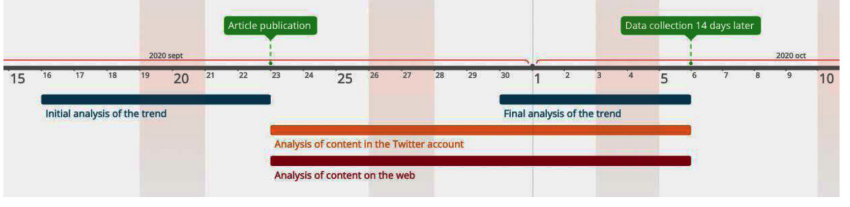



**Figure 3**

**Timeline of data collection from an article published on 23 September 2020.**

and the success of these themes in the time interval studied. These key concepts can be of the following types, framed within the subject of cinema being analysed in this study:

- Film titles, preferably in the language of the medium and in the original version if this is unavailable.
- Names of entities, such as directors, actors, scriptwriters, production companies, distributors, etc.
- Names of events, preferably in the language of the medium and in the original version if not available, such as film festivals.

It is necessary to remember that these key concepts must be directly related to the content, not circumstantial or contextual. This selection of terms has been carried out manually by the editor of each article in the digital media.

The following methodology is used to study the possible prediction of the variables whose data have been obtained, as the correlation, variability, and possible calculation of these variables are analysed using a multiple linear regression based on the variables that are counted at the time of publication of an article.

This means that variables or success scenarios are established as those that may be of interest to the editorial team of a medium, such as detecting which trends tend to have a more significant number of tweets 14 days later (they remain stable or grow over time), which articles tend to have a lower percentage of exit, a more significant number of unique page views or which articles elicit a better response in the form of retweets or favourites in the medium's own Twitter account. And since editorial decisions are made before the article is published, only the data related to the trend analysis carried out on the day of publication is available.

**Determine which statistical procedures, tools, and reports are necessary.**

A statistical methodology based on machine learning has been carried out, divided into two phases: a data training phase to obtain the proposed prediction calculations and a data testing phase to check the validity of the model that has been generated. The data set is usually divided into 70% training and 30% test data [43]. In this way, there is sufficient training data to make statistical models with which to extract, as in this case, prediction equations. Subsequently, these equations will be validated with the test

data, thus extracting their accuracy since they will try to make their predictions with new and different data from those used for their creation.

A first phase of training data collection has been carried out using the information about the breaking news published in two calendar months, between 23 September and 22 November 2020, and the related news published previously in the same digital media has been captured.

The following process has been followed for the sets of all articles, film articles, and trailer articles:

(1) Success variables.
   (a) Analyse each variable.
   (b) Assess the normality and equidistribution of the residuals.
   (c) Normalise variables if necessary.
   (d) Filter out strong correlations.
(2) Predictor variables.
   (a) Analyse each variable.
   (b) Assess normality, equal distribution of residuals, and homocedasticity.
   (c) Normalise variables if necessary.
   (d) Filter out strong correlations.
   (e) Perform principal component analysis.

Once these preliminary analyses were performed, a list of the success or dependent variables and the predictor or independent variables that met the requirements to be used in a multiple linear regression was obtained.

*Success or dependent variables*

In the case of this study, success variables were defined as those that allowed a valid multiple linear regression model to be extracted from the existing data:

- Unique page views (total): unique_page_visits_sum.
- Duration of visit (average): avg_time_on_page_mean.
- Pageviews per session (average): avg_pageviews_per_session.
- Number of favourites in the media account (average): mean_favorite_count.

- Number of retweets of the trend 14 days later (average): mean_retweets_after_14d.

*Predictive or independent variables*

At the time of the article's publication, the following variables are available, which are the only ones we can select as predictor variables:

- Number of tweets of the initial trend (total): initial_trend_tweet_count.
- Number of retweets of the initial trend (total): initial_trend_retweets_sum.
- Number of retweets of the initial trend (average): initial_trend_retweets_mean.
- Number of favourites of the initial trend (total): initial_trend_favorites_sum.
- Number of favourites of the initial trend (average): initial_trend_favorites_mean.
- Number of medium followers talking about the initial trend (total): initial_trend_followers_discussion_count.
- Number of followers of users talking about the initial trend (average): initial_trend_user_followers_avg.
- Number of tweets of users talking about the initial trend (average): initial_trend_user_tweets_avg.
- Age in number of days of the accounts of the users who speak of the initial trend (average): initial_trend_user_account_age_avg.
- Ratio of inclusion of URLs in the tweets of the initial trend: initial_trend_url_ratio.

**Use regression analyses to develop predictive models for the chosen measures of success.**

After analysing the indicators' variability and correlation, a series of equations have been proposed to predict the success variables chosen from the initial set. The analysis consists of the selection and weighting of the indicators through multiple regressions, provided that the model has been significant, i.e., the predictive variables available can predict the selected success variable. For this to be possible, certain requirements must be met: linearity, normality and equal distribution of residuals,

avoidance of collinearity, and identification and discarding of outliers where appropriate. On the other hand, homoscedasticity is a desirable property in a linear regression model, as it allows for a more reliable model. It means that the variance of the errors is constant over time [34], so attention will also be paid to this characteristic of the variables in the model.

**Evaluate the predictive models with test data to determine their accuracy and effectiveness.**

The analysis of the results has been carried out by taking as a universe the news set published in a digital medium. The digital medium hellofriki.com has been accepted as a use case due to the availability of access to all the information necessary to carry out this study. It is a website with informative online entertainment content, with editorial lines on cinema, series, video games, literature, and comics. It has been active since 2009 and has published over ten thousand articles. In 2020, the year in which the data for this study was taken, it had an annual traffic of more than 210,000 users and almost 315,000 page views, so it serves as a sample of the population of websites that publish cinema news. This work can also be seen as a testing ground for similar areas.

As an object of study, we have considered the cinema news published during three calendar months in this medium, between 23 September and 22 December 2020. All articles published in that period and trailer news were also considered, as they are thematically related to cinema and are the most topical subpopulations with a sufficient number of articles to draw statistical conclusions, as the minimum number of articles to be able to draw valid conclusions is 30 [41].

After obtaining the equations for the success variables, the second phase was carried out, consisting of validation in one calendar month, between 23 November 2020 and 22 December 2020. This way, the methodology proposed in this study is validated by comparing the predictions with the values obtained this month.

This comparison is made by calculating the coefficient of determination (R-squared) [33], which represents the portion of a variable's overall variance that is accounted for by the regression model. It is the foundation for the predictive equations developed in the initial phase [50].

The objective is to validate the prediction equations obtained in the regressions. For this purpose, we will utilise the Root Mean Square Error (RMSE), essentially the standard deviation of errors of prediction, to

quantify how far observed data points deviate from the fitted regression line. Therefore, the root mean squared error measures the extent to which the residuals are dispersed or, in other words, the concentration or closeness of the data to the line of best fit defined by the equation [17]. The lower the RMSE value, the better the fit of the prediction to the data. RMSE is an appropriate measure of the accuracy with which the model predicts the data, and the most important criterion if the objective is the prediction itself [18].

With this, it will be possible to check to what extent the prediction has been wrong according to the available equations and, therefore, to what extent its forecast impacts the process of shaping the publication's editorial strategy. The phase 1 RMSE, for which the prediction equations were obtained, and the phase 2 RMSE with a new (test) data set will be compared. If both are similar, the validity of the prediction is confirmed.

On the other hand, to also check the accuracy, the Scatter Index (SI) will also be obtained, which consists of normalising the RMSE by the mean value of the observed data. According to Campos and Soares (2016), the lower the value of the RMSE and especially the Scatter Index (since this is calculated relative to each variable), the more accurate the prediction will be [10].

## Results

**Phase 1:** Analysis of training data

In phase 1, which lasted two months, 350 articles were published, of which 187 were about films and 101 were about trailers. A total of 7,035 related articles and 817 related terms were analysed.

**Phase 1.1:** Multiple linear regressions with data from all articles

After having performed all the analyses before the multiple linear regressions, Table 4 shows the list of predictor or dependent variables and success or independent variables.

**Table 4**

**All items: Final list of predictor and success variables for multiple linear regression.**

| Predictor variables | Success variables |
|---|---|
| log(initial_trend_retweets_mean) | log(unique_page_visits_sum) |

*Contd...*

| | |
|---|---|
| log(initial_trend_favorites_mean) | log(avg_time_on_page_mean) |
| log(initial_trend_user_followers_avg) | log(avg_pageviews_per_session) |
| log(initial_trend_url_ratio) | log(mean_favorite_count) |
| | log(mean_retweets_after_14d) |

After performing the multiple linear regressions relevant to each success variable, the following models have been obtained and simplified by eliminating any variables whose P-value exceeds 0.05. This confirms that the relationship among the retained variables is statistically significant, with 95% confidence.

*Unique pages (total).*

The formula of the fitted model is:

unique_page_visits_sum = exp(3.85 + 0.07 * log(initial_trend_retweets_mean) – 0.1 * log(initial_trend_user_followers_avg))

According to the R-Squared, the model thus adjusted explains 2.88% of the variability of log(unique_page_visits_sum), while the adjusted R-Squared indicates 2.31%.

*Duration of visit (average).*

The formula of the fitted model is:

avg_time_on_page_mean = exp(4.16 – 0.29 * log(initial_trend_url_ratio))

According to the R-Squared, the model thus fitted explains 1.61% of the variability of log(avg_time_on_page_mean), while the adjusted R-Squared indicates 1.32%.

*Page views per session (average).*

The formula for the fitted model is:

avg_pageviews_per_session = exp(0.35 – 0.11 * log(initial_trend_favorites_mean))

According to the R-Squared, the model thus adjusted explains 2.24% of the variability of log(avg_pageviews_per_session), while the adjusted R-Squared indicates 1.95%.

*Number of favourites in the media account (average).*

The formula for the fitted model is:

mean_favorite_count = exp(0.18 – 0.15 * log(initial_trend_url_ratio))

According to the R-Squared, the model thus fitted explains 1.63% of the variability of log(favourite_count_mean), while the adjusted R-Squared indicates 1.31%.

*Number of retweets of the trend 14 days later (average).*

The formula for the fitted model is:

mean_retweets_after_14d = exp(0.83 + 0.63 * log(initial_trend_retweets_mean))

According to the R-Squared, the model thus adjusted explains 33.34% of the variability of log(mean_retweets_after_14d), while the adjusted R-Squared indicates 33.14%.

**Phase 1.2:** Multiple linear regressions on film article data

After having carried out all the analyses before the multiple linear regressions, Table 5 shows the list of predictor or dependent variables and success or independent variables.

**Table 5**

**Film articles: Final list of predictor and hit variables for multiple linear regression.**

| Predictor variables | Success variables |
|---|---|
| log(initial_trend_retweets_mean) | log(unique_page_visits_sum) |
| log(initial_trend_favorites_mean) | log(avg_time_on_page_mean) |
| log(initial_trend_followers_discussion_count) | log(avg_pageviews_per_session) |
| log(initial_trend_user_followers_avg) | log(mean_favorite_count) |
| log(initial_trend_url_ratio) | log(mean_retweets_after_14d) |

After performing the multiple linear regressions relevant to each success variable, the following models have been obtained and simplified by eliminating any variables whose P-value exceeds 0.05. This confirms that the relationship among the retained variables is statistically significant, with 95% confidence.

*Unique pages (total).*

The formula of the fitted model is:

unique_page_visits_sum = exp(2.25 – 0.19 * log(initial_trend_followers_discussion_count))

According to the R-Squared, the model thus adjusted explains 3.26% of the variability of log(unique_page_visits_sum), while the adjusted R-Squared indicates 2.65%.

*Duration of visit (average).*

The formula for the fitted model is:

avg_time_on_page_mean = exp(6.1 – 0.18 * log (initial_trend_user_followers_avg))

According to the R-Squared, the model thus adjusted explains 4.02% of the variability of log(avg_time_on_page_mean), while the adjusted R-Squared indicates 3.5%.

*Page views per session (average).*

The formula for the fitted model is:

avg_pageviews_per_session = exp(1.29 + 0.09 * log(initial_trend_followers_discussion_count) – 0.07 * log(initial_trend_user_followers_avg))

According to the R-Squared, the model thus adjusted explains 5.86% of the variability of log(avg_pageviews_per_session), while the adjusted R-Squared indicates 4.66%.

*Number of favourites in the media account (average).*

The formula for the fitted model is:

mean_favorite_count = exp(-0.22 – 0.12 * log (initial_trend_followers_discussion_count))

According to the R-Squared, the model thus fitted explains 3.89% of the variability of log(favourite_count_mean), while the adjusted R-Squared indicates 3.16%.

*Number of retweets of the trend 14 days later (average).*

The formula for the fitted model is:

mean_retweets_after_14d = exp(-0.68 + 0.47 * log(initial_trend_retweets_mean) – 0.52 * log(initial_trend_followers_discussion_count))

According to the R-Squared, the model thus adjusted explains 33.34% of the variability of log(mean_retweets_after_14d), while the adjusted R-Squared indicates 32.45%.

**Phase 1.3:** Multiple linear regressions with trailer item data

After having performed all the analyses before the multiple linear regressions, Table 6 shows the list of predictor or dependent variables and success or independent variables.

**Table 6**

**Trailer items: Final list of predictor and hit variables for multiple linear regression.**

| Predictor variables | Success variables |
|---|---|
| log(initial_trend_tweet_count) | log(unique_page_visits_sum) |
| log(initial_trend_retweets_mean) | log(avg_time_on_page_mean) |
| log(initial_trend_favorites_mean) | log(avg_pageviews_per_session) |
| log(initial_trend_followers_discussion_count) | log(mean_favorite_count) |
| log(initial_trend_user_followers_avg) | log(mean_retweets_after_14d) |
| log(initial_trend_user_tweets_avg) | |
| log(initial_trend_url_ratio) | |

After performing the multiple linear regressions relevant to each success variable, the following models have been obtained and simplified by eliminating any variables whose P-value exceeds 0.05. This confirms that the relationship among the retained variables is statistically significant, with 95% confidence.

*Unique pages (total).*

The formula of the fitted model is:

$$\text{unique_page_visits_sum} = \exp(2.23 + 0.1 * \log(\text{initial_trend_tweet_count}))$$

According to the R-Squared, the model thus adjusted explains 6.04% of the variability of log(unique_page_visits_sum), while the adjusted R-Squared indicates 5.09%.

*Duration of visit (average).*

The formula for the fitted model is:

avg_time_on_page_mean = exp(3.84 – 0.44 * log(initial_trend_favorites_mean) – 0.35 * log(initial_trend_followers_discussion_count))

According to the R-Squared, the model thus adjusted explains 10.94% of the variability of log(avg_time_on_page_mean), while the adjusted R-Squared indicates 8.72%.

*Page views per session (average).*

The formula for the fitted model is:

avg_pageviews_per_session = exp(3.83 + 0.18 * log(initial_trend_followers_discussion_count) – 0.27 * log(initial_trend_user_tweets_avg))

According to the R-Squared, the model thus adjusted explains 18.35% of the variability of log(avg_pageviews_per_session), while the adjusted R-Squared indicates 16.31%.

*Number of favourites in the media account (average).*

The formula for the fitted model is:

mean_favorite_count = exp(-0.5 + 0.14 * log(initial_trend_tweet_count) – 0.18 * log(initial_trend_favorites_mean))

According to the R-Squared, the model thus adjusted explains 18.26% of the variability of log(favourite_count_mean), while the adjusted R-Squared indicates 16.17%.

*Number of retweets of the trend 14 days later (average).*

The formula for the fitted model is:

mean_retweets_after_14d = exp(-1.86 + 0.49 * log(initial_trend_tweet_count) + 0.34 * log(initial_trend_retweets_mean))

According to the R-Squared, the model thus adjusted explains 45.23% of the variability of log(mean_retweets_after_14d), while the adjusted R-Squared indicates 44.02%.

*Summary of model results for all variables*

After performing all the above multiple linear regressions, the R-Squared of each model explains the variability explained by the models.

Table 7 shows the R-Squared of the three existing models for each variable analysed in this article.

**Table 7**

**R-Squared summary of all multiple linear regression models for each variable.**

| Variable | R-Squared All | R-Squared Cinema | R-Squared Trailers |
|---|---|---|---|
| unique_page_visits_sum | 2.88% | 3.26% | 6.04% |
| avg_time_on_page_mean | 1.61% | 4.02% | 10.94% |
| avg_pageviews_per_session | 2.24% | 5.86% | 18.35% |
| mean_favorite_count | 1.63% | 3.89% | 18.26% |
| mean_retweets_after_14d | 33.34% | 33.34% | 45.23% |

In Table 7, it can be observed that the variability explained by the models increases for more specific sub-populations in all cases.

**Phase 2:** Analysis of test data

In phase 2, which lasted one month, 178 articles were published, of which 80 were about films, and 57 were about trailers. A total of 5,251 related articles and 462 related terms were analysed.

Table 8 shows the R-Squared of the prediction of each type of regression, the root mean squared error (RMSE), and the Scatter Index (SI) resulting from comparing the predicted values with those observed in all the articles published in both phase 1 and phase 2.

**Table 8**

**Measured errors for each success variable according to Multiple Linear Regression considering all articles.**

| Variable | RMSE (phase 1) | RMSE (phase 2) | SI (phase 2) |
|---|---|---|---|
| unique_page_visits_sum | 64.54 | 306.93 | 4.05 |
| avg_time_on_page_mean | 140.43 | 105.15 | 1.07 |
| avg_pageviews_per_session | 0.78 | 0.82 | 0.6 |
| mean_favorite_count | 0.99 | 1.02 | 0.99 |
| mean_retweets_after_14d | 276.88 | 152.16 | 1.89 |

Total pageviews (unique_page_visits_sum) had a significantly higher RMSE in phase 2, leading to a huge SI of more than 4. The RMSE of pageviews per session (avg_pageviews_per_session) and the average number of favourites in the media account (favourite_count_mean) have similar values and a lower SI than the other variables. The average duration of the visit (avg_time_on_page_mean) and the average number of retweets of the trend (mean_retweets_after_14d) have decreased significantly but with an SI greater than 1.

**Table 9**

**Measured errors for each success variable according to Multiple Linear Regression considering film articles.**

| Variable | RMSE (phase 1) | RMSE (phase 2) | SI (phase 2) |
|---|---|---|---|
| unique_page_visits_sum | 38.54 | 96.06 | 1.27 |
| avg_time_on_page_mean | 133.54 | 88.79 | 0.9 |
| avg_pageviews_per_session | 0.85 | 0.94 | 0.69 |
| mean_favorite_count | 1.09 | 1 | 0.98 |
| mean_retweets_after_14d | 273.66 | 88.78 | 1.1 |

As shown in Table 9, the value of the SI of the predictions with data from film articles is generally lower than in the predictions, taking into account all articles. This shows that, except for the average number of page views per session (avg_pageviews_per_session) by a difference of 0.09, the prediction with data from film articles is better than data from all articles.

**Table 10**

**Measured errors for each success variable according to Multiple Linear Regression considering trailer items.**

| Variable | RMSE (phase 1) | RMSE (phase 2) | SI (phase 2) |
|---|---|---|---|
| unique_page_visits_sum | 70.08 | 57.6 | 0.76 |
| avg_time_on_page_mean | 150.51 | 150.44 | 1.53 |
| avg_pageviews_per_session | 1 | 1.2 | 0.88 |
| mean_favorite_count | 1.21 | 1.01 | 0.98 |
| mean_retweets_after_14d | 154.9 | 133.78 | 1.66 |

Table 10 shows that the pageviews (unique_page_visits_sum) of trailer articles have a much lower SI than all articles and cinema articles. It slightly outperformed the set of all articles on the average number of

favourites in the media account (favourite_count_mean) and the average number of retweets of the trend (mean_retweets_after_14d). It had a higher SI than the cinema subset on all variables except pageviews.

### Selection of prediction equations

The lowest SI will be used as a scale to select the prediction equations for each variable and each section or thematic selection. To select the prediction formula for a success variable of interest, one must choose the one with the lowest SI among its different possibilities. For example:

- If you want to predict the average pageview time (avg_time_on_page_mean) of a Cinema news item, you would go for the prediction formula of the subset Cinema, with an SI of 0.9 lower than the SI of all articles of 1.07.
- If one chooses to predict the pageviews (unique_page_visits_sum) of a Cinema trailer news item, one would use the prediction formula of the subset Trailers, with an SI of 0.75 less than the SI of all articles of 4.05 and the SI of the subset Cinema of 1.27.

## Discussion and conclusion

This work has achieved the main objective of designing a methodology to predict an online post success related to cinema based on the data available from web analytics and the analysis of trends related to the publication's content.

The first objective was to select the most appropriate cybermetric indicators. The selected indicators were divided into three groups: web content analytics, content analytics on the media outlet's Twitter account, and Twitter trend analysis. Thus, each group responded to a different dimension of success. Twitter, in particular, is characterised by the real-time publication of huge amounts of content from its users, so any approach that takes this into account faces challenges such as dealing effectively with informal language, detecting events without having a predefined number of events, and scaling to the large volume of Twitter data [28].

Before data extraction and processing, the second objective focused on selecting an example that could be extrapolated to the population under study. In this article, the digital medium hellofriki.com has been taken as a use case because it has a sufficient number of articles to draw

statistical conclusions and because of the availability of access to all the information, which has made this study possible.

The third objective defined indicators as success or independent variables, predictor or dependent variables, and data collection. Tools such as the Google Analytics v4 reporting API and the Twitter Standard API made it possible to obtain the data necessary for the study. With this, preliminary analyses have been carried out to satisfy the conditions for multiple linear regressions, as indicated by authors such as [6], [9], and [37].

The fourth objective was based on the initial design of the methodology and the determination of the statistical procedures necessary to determine which predictor (independent) and success (dependent) variables can be used. Thanks to this step, it has been possible to fulfil the fifth objective by carrying out the multiple linear regressions that have made it possible to obtain the prediction equations for the success variables. Thus, it has been determined that the indicators proposed by authors such as [27], [20], [51], and [53] have indeed provided part of the variability necessary to explain the success indicators. However, in this work, a relatively low R-Squared in some cases has been remarkable, which has led to a decrease in accuracy. This means that if a prediction formula has an R-Squared of 18.35%, as in the case of the avg_pageviews_per_session variable of Trailers, there is a variability of 81.65% that is not taken into account and which can significantly affect the accuracy of the model. A more significant number of indicators could represent a wide variability for different interpretations of success depending on the context.

The validation of these prediction equations with test data has been the sixth objective of this work. It has made it possible to establish the degree of confidence to be placed in each prediction formula and establish a basis for improving it in future research.

This paper has not only demonstrated that some of the information in the success indicators can be predicted but also encourages more granular and specific analyses, as an improvement in the variability explained (R-squared) by the models has been observed in more particular sub-populations such as trailer articles. To improve the accuracy of the predictions, one could choose to analyse much longer time intervals, examine outliers separately from the rest of the data, try to find new predictor variables that can increase the explained variability, and select much more specific thematic units, provided that there is a sufficient population of articles to generate statistical conclusions.

**Declaration of Interest Statement**

No conflicts of interest were disclosed by the author(s).

**Data Availability Statement**

The data underlying this study's results are accessible in Figshare[1].

**References**

[1] W. Ahmed and S. Lugovic, "Social media analytics: analysis and visualisation of news diffusion using NodeXL", *Online Information Review*, Vol. 43 No. 1, pp. 149-160 (2019). https://doi.org/10.1108/OIR-03-2018-0093.

[2] A. bin A. Al-Tameem and A. K. J. Saudagar, "Machine learning approach for identification of threat content in audio messages shared on social media", *Journal of Discrete Mathematical Sciences and Cryptography*, Vol. 23 No. 1, pp. 83–93 (2020). https://doi.org/10.1080/09720529.2020.1721876.

[3] V. Álvarez Intriago, L. Agreda Fernández, and A. Cevallos Gamboa, "Análisis de la estrategia de marketing digital mediante herramientas de analítica web", *INVESTIGATIO*, Vol. *7*, pp. 81–97 (2016). https://doi.org/10.31095/IRR.V0I7.41.

[4] T. Araujo and J. Kollat, "Communicating effectively about CSR on Twitter: The power of engaging strategies and storytelling elements", *Internet Research*, Vol. *28* No. 2, pp. 419-431 (2018). https://doi.org/10.1108/IntR-04-2017-0172.

[5] S. Asur, B. A. Huberman, G. Szabo, and C. Wang, "Trends in Social Media: Persistence and Decay", *SSRN Electronic Journal* (2011). https://doi.org/10.2139/ssrn.1755748.

[6] F. J. Barón López and F. Téllez Montiel, "Apuntes de Bioestadística: Tercer Ciclo en Ciencias de la Salud y Medicina", *Apuntes de Bioestadística*, Málaga (2004). https://www.bioestadistica.uma.es/baron/apuntes/ficheros/cap06.pdf.

[7] V. Belair-Gagnon and A. E. Holton, "Boundary Work, Interloper Media, And Analytics In Newsrooms: An analysis of the roles of web analytics companies in news production", *Digital Journalism*, Vol. *6*

[1] https://doi.org/10.6084/m9.figshare.16553061.v2

No. 4, pp.492–508 (2018). https://doi.org/10.1080/21670811.2018.1445001.

[8] S. Bhoyar, P. Bhoyar, and M. A. Shah, A machine learning-based predictive approach in evaluating consumer behavior, *Journal of Statistics and Management Systems*, Vol. 26 No. 8, 1955–1963 (2023). https://doi.org/10.47974/JSMS-1131.

[9] Z. Bobbitt, "A Gentle Introduction to Poisson Regression for Count Data" (2019), available at: https://www.statology.org/poisson-regression/ (accessed 20 December 2020).

[10] R. M. Campos and C. G. Soares, "Comparison and assessment of three wave hindcasts in the North Atlantic Ocean", *Journal of Operational Oceanography*, Vol. *9* No. 1, pp. 26–44 (2016). https://doi.org/10.1080/1755876X.2016.1200249.

[11] M. Carlson, "Confronting Measurable Journalism", *Digital Journalism*, Vol. *6* No. 4, pp. 406–417 (2018). https://doi.org/10.1080/21670811.2018.1445003.

[12] C. Castillo, M. Mendoza, and B. Poblete, "Predicting information credibility in time-sensitive social media", *Internet Research*, Vol. *23* No. 5, 560-588 (2013). https://doi.org/10.1108/IntR-05-2012-0095.

[13] M. Cristian, P. Morphl, and A. A. Morphl, "Predicting next shopping stage using Google Analytics data for E-commerce applications" (2019), unpublished manuscript, arXiv preprint, arXiv:1905.12595

[14] J. Díaz Noci, "Definición teórica de las características del ciberperiodismo: elementos de la comunicación digital", *Doxa Comunicación. Revista Interdisciplinar de Estudios de Comunicación y Ciencias Sociales*, Vol. *6*, pp. 53–91 (2008). https://doi.org/10.31921/doxacom.n6a2.

[15] e-intelligent.es., "Medir el éxito de nuestro Marketing de Contenidos (2014)", available at: https://www.e-intelligent.es/es/blog/medir-el-exito-de-nuestro-marketing-de-contenidos (accessed 21 May 2019).

[16] S. Fox and A. Lenhart, "Twitter and status updating," (2009). [Online]. Available: http://www.pewinternet.org/Reports/2009/Twitter-and. [Accessed: Jan. 11, 2020].

[17] S. Glen, "RMSE: Root Mean Square Error," (2020). [Online]. Available: https://www.statisticshowto.com/probability-and-statistics/regression-analysis/rmse-root-mean-square-error/. [Accessed: Dec. 24, 2020].

[18] K. Grace-Martin, "Assessing the Fit of Regression Models - The Analysis Factor," (2008). [Online]. Available: https://www.theanalysisfactor.

[19] A. Graefe, "Guide to Automated Journalism," *Tow Center for Digital Journalism* (2016). [Online]. Available: https://doi.org/10.7916/D80G3XDJ.

[20] C. Gutiérrez Argüello, "Modelos de Comunicación y Negocios en Periódicos Digitales," (2013). [Online]. Available: https://es.slideshare.net/CarlosGutArg/modelos-de-comunicacin-y-negocios-en-peridicosdigitales. [Accessed: May 6, 2019].

[21] A. Hermida, "Social Media and the News," in *The SAGE Handbook of Digital Journalism*, SAGE Publications Ltd., London, pp. 81–94 (2016).

[22] L. Hong, O. Dan, and B. D. Davison, "Predicting Popular Messages in Twitter," in Proceedings of the 20th International Conference Companion on World Wide Web, pp. 57-58 (2011). [Online]. Available: http://bit.ly/794Edz.

[23] J. Huang, Y. Tang, Y. Hu, J. Li, and C. Hu, "Predicting the active period of popularity evolution: A case study on Twitter hashtags", *Information Sciences*, Vol. 512, pp.315–326 (2020). https://doi.org/10.1016/j.ins.2019.04.028.

[24] A. Java, X. Song, T. Finin, and B. Tseng, "Why We Twitter: An analysis of a microblogging community," in International Workshop on Social Network Mining and Analysis, pp. 118-138 (2009). [Online]. Available: https://doi.org/10.1007/978-3-642-00528-2_7.

[25] V. Karnowski, D. J. Leiner, A. S. Kümpel, and L. Leonhard, "Worth to Share? How Content Characteristics and Article Competitiveness Influence News Sharing on Social Network Sites", *Journalism & Mass Communication Quarterly*, Vol. *98* No. 1, pp. 59–82 (2021). https://doi.org/10.1177/1077699020940340.

[26] A. Kaushik, "Best Social Media Metrics: Conversation, Amplification, Applause, Economic Value," (2011). [Online]. Available: https://www.kaushik.net/avinash/best-social-media-metrics-conversation-amplification-applause-economic-value/. [Accessed: May 21, 2019].

[27] A. Kaushik, "Best Web Metrics / KPIs for a Small, Medium or Large Sized Business," (2011). [Online]. Available: https://www.kaushik.net/avinash/best-web-metrics-kpis-small-medium-large-business/. [Accessed: Feb. 5, 2020].

[28] S. Kumar, H. Liu, S. Mehta, and L. V. Subramaniam, "From Tweets to Events: Exploring a Scalable Solution for Twitter Streams," unpublished manuscript, arXiv preprint, arXiv:1405.1392 (2014).

[29] H. Kwak, C. Lee, H. Park, and S. Moon, "What is Twitter, a social network or a news media?" in Proceedings of the 19th International Conference on World Wide Web, pp. 591-600 (2010).

[30] T. Laor, "Twitter as a clique: journalists' patterns of Twitter use in Israel", *Online Information Review*, Vol. 46 No. 1, pp. 40-58 (2022). https://doi.org/10.1108/OIR-07-2020-0324.

[31] J. Li and C. Cardie, "Timeline Generation: Tracking individuals on Twitter," in Proceedings of the 23rd International Conference on World Wide Web, pp. 643-652 (2014).

[32] Z. Liu, L. Liu, and H. Li, "Determinants of information retweeting in microblogging", *Internet Research*, Vol. *22* No. 4, pp. 443-466 (2012). https://doi.org/10.1108/10662241211250980.

[33] J. F. López, "Coeficiente de determinación (R cuadrado) - Qué es, definición y concepto," (2017). [Online]. Available: https://economipedia.com/definiciones/r-cuadrado-coeficiente-determinacion.html. [Accessed: Nov. 22, 2020].

[34] J. F. López, "Homocedasticidad - Qué es, definición y concepto," (2017). [Online]. Available: https://economipedia.com/definiciones/homocedasticidad.html. [Accessed: Dec. 18, 2020].

[35] P. Malhotra and S. K. Malik, "Fake news detection using supervised machine learning techniques", *Journal of Information and Optimization Sciences*, Vol. 43 No. 1, pp. 7–15 (2022). https://doi.org/10.1080/02522667.2022.2038933.

[36] M. Muñoz-Expósito, M. Á. Oviedo-García, and M. Castellanos-Verdugo, "How to measure engagement in Twitter: Advancing a metric," *Internet Research,* vol. 27, no. 5, pp. 1122-1148 (2017). [Online]. Available: https://doi.org/10.1108/IntR-06-2016-0170.

[37] A. Navarro, F. Utzet, P. Puig, J. Caminal, and M. Martín, "Negative binomial distribution versus Poisson in the analysis of recurrent phenomena," Gaceta Sanitaria / S.E.S.P.A.S, vol. 15, no. 5, pp. 447-452 (2001). [Online]. Available: https://doi.org/10.1016/S0213-9111(01)71599-3.

[38] J. M. Noguera, "La convergencia entre cibermedios y blogosfera: apuntes para entrar de forma natural en la conversación", *Temas de Comunicación*, Vol. 18, pp.135–151 (2009).

[39] B. Omar, H. Al-Samarraie, and B. Wright, "Immediacy as news experience: exploring its multiple dimensions in print and online contexts", *Online Information Review*, Vol. 45 No. 2, pp. 461-480 (2021). https://doi.org/10.1108/OIR-12-2019-0388.

[40] C. Orellana-Rodriguez and M. T. Keane, "Attention to news and its dissemination on Twitter: A survey", *Computer Science Review*, Vol. *29*, pp. 74–94 (2018). https://doi.org/10.1016/j.cosrev.2018.07.001.

[41] D. Peña, *Fundamentos de estadística*, Alianza editorial, Madrid (2014).

[42] C. Peñafiel Sáiz, "Reinvención del periodismo en el ecosistema digital y narrativas transmedia", *AdComunica*, Vol. 12, pp.163–182 (2016).

[43] P. Recuero de los Santos, "Datos de entrenamiento vs datos de test," (2020). [Online]. Available: https://empresas.blogthinkbig.com/datos-entrenamiento-vs-datos-de-test/. [Accessed: Nov. 22, 2020].

[44] L. Rodríguez Martínez, "5 indicadores clave para medir la interacción en las redes sociales," (2013). [Online]. Available: https://www.puromarketing.com/42/16086/indicadores-clave-para-medir-interaccionredes-sociales.html. [Accessed: May 21, 2019].

[45] W. Ryan, W. R. Hazlewood, and K. Makice, "Twitterspace: A co-developed display using Twitter to enhance community awareness," in Proceedings of the Tenth Conference on Participatory Design, pp. 230–233 (2008). [Online]. Available: https://doi.org/10.1145/1795234.1795284.

[46] N. Savage, "Twitter as medium and message", *Communications of the ACM*, Vol. 54, No. 3, pp. 18–20 (2011). https://doi.org/10.1145/1897852.1897860

[47] M. Schudson, "The crisis in news: Can you whistle a happy tune?" in The Crisis of Journalism Reconsidered: Democratic Culture, Professional Codes, Digital Future, Cambridge University Press, pp. 98–115 (2016).

[48] C. M. Segijn, E. Maslowska, T. Araujo, and V. Viswanathan, "Engaging with TV events on Twitter: The interrelations between TV consumption, engagement actors, and engagement content", *Internet Research*, Vol. 30 No. 2, pp. 381-401 (2019). https://doi.org/10.1108/INTR-08-2018-0389

[49] A. Shiri, "Cybermetrics: A New Horizon in Information Research", *Cybermetrics*, Vol. 50, p. 111 (1998).

[50] R. G. D. Steel and J. H. Torrie, *Principles and procedures of statistics*, McGraw-Hill Book Company, Inc, New York, Toronto, London (1960).

[51] B. Suh, L. Hong, P. Pirolli, and E. H. Chi, "Want to be Retweeted? Large Scale Analytics on Factors Impacting Retweet in Twitter Network", *2010 IEEE Second International Conference on Social Computing*, pp. 177–184 (2010). https://doi.org/10.1109/SocialCom.2010.33.

[52] J. Teevan, D. Ramage, and M. R. Morris, "#TwitterSearch: A Comparison of Microblog Search and Web Search", *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, pp. 35–44 (2011).

[53] M. Thelwall and B. Cugelman, "Monitoring Twitter strategies to discover resonating topics: The case of the UNDP," El Profesional de La Información, vol. 26, no. 4 (2017). [Online]. Available: https://doi.org/10.3145/epi.2017.jul.09.

[54] Twitter Developers, "GET followers/ids," (2020). [Online]. Available: https://developer.twitter.com/en/docs/accounts-and-users/follow-search-get-users/api-reference/get-followers-ids. [Accessed: Aug. 7, 2020].

[55] Twitter Developers, "GET statuses/user_timeline," (2020). [Online]. Available: https://developer.twitter.com/en/docs/tweets/timelines/api-reference/get-statuses-user_timeline. [Accessed: Jun. 30, 2020].

[56] Twitter Developers, "Pricing," (2020). [Online]. Available: https://developer.twitter.com/en/pricing. [Accessed: Jun. 23, 2020].

[57] Twitter Developers, "Rate limiting," (2020). [Online]. Available: https://developer.twitter.com/en/docs/basics/rate-limiting. [Accessed: Aug. 7, 2020].

[58] Twitter Developers, "Standard search API," (2020). [Online]. Available: https://developer.twitter.com/en/docs/tweets/search/api-reference/get-search-tweets. [Accessed: Jun. 30, 2020].

[59] Twitter Developers, "Things every developer should know," (2020). [Online]. Available: https://developer.twitter.com/en/docs/basics/thingsevery-developer-should-know. [Accessed: Aug. 7, 2020].

[60] J. Yang and S. Counts, "Predicting the Speed, Scale, and Range of Information Diffusion in Twitter", *Fourth International AAAI Conference on Weblogs and Social Media*, pp. 355-358 (2010). https://doi.org/10.1609/icwsm.v4i1.14039.

[61] A. Zubiaga, D. Spina, R. Martínez, and V. Fresno, "Real-time classification of Twitter trends", *Journal of the Association for Information Science and Technology*, Vol. *66* No. 3, pp. 462–473 (2015). https://doi.org/10.1002/asi.23186

[62] D. Zumstein and M. Kaufmann, "A fuzzy Web analytics model for Web mining", *Proceedings of Data Mining 2009*, pp.59–66 (2009).

*Received December, 2023*